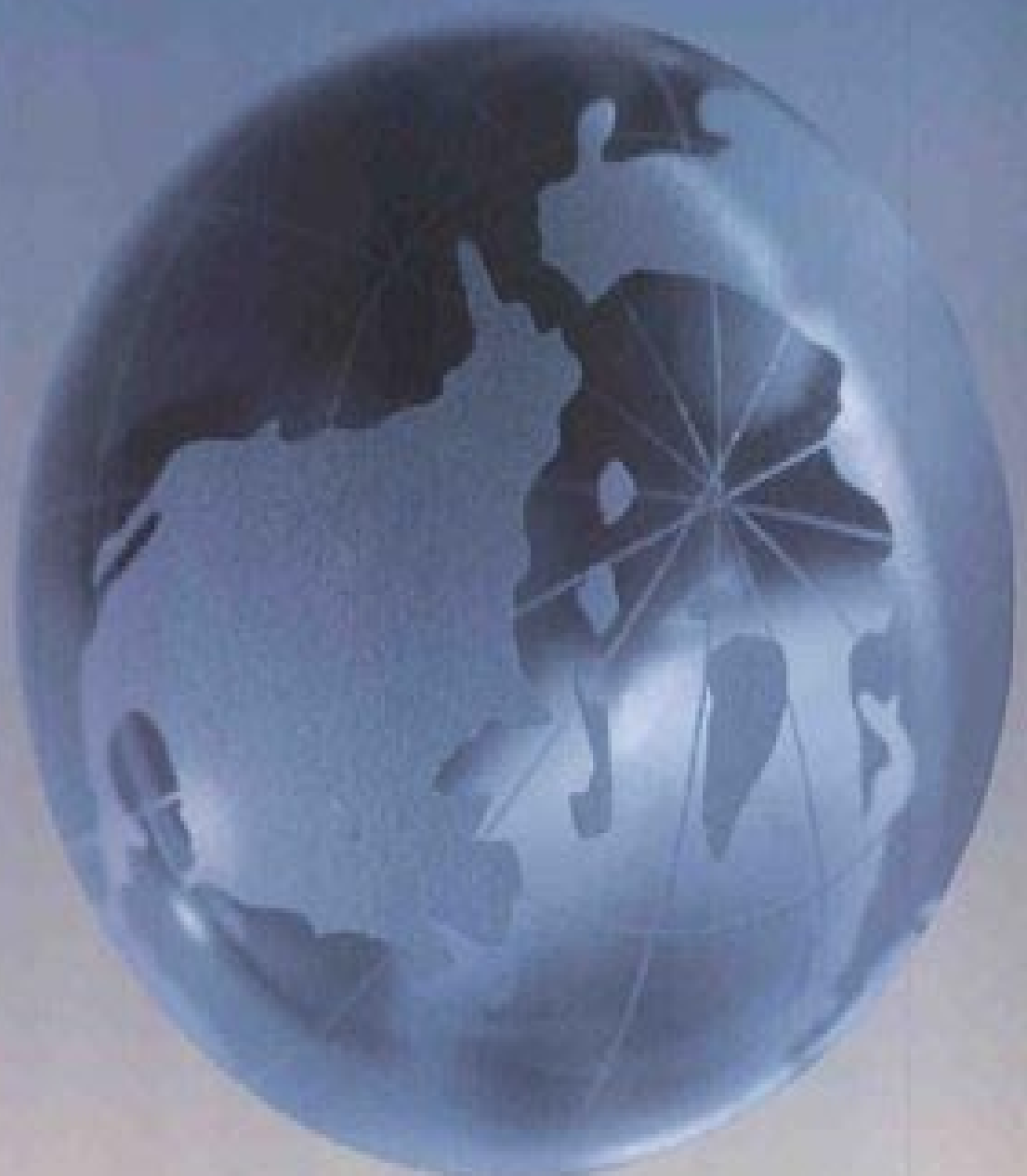


The Professional Risk Managers' Handbook

A Comprehensive Guide to Current Theory and Best Practices

Edited by Carol Alexander
and Elizabeth Sheedy



The Professional Risk Managers' Handbook

A Comprehensive Guide to Current Theory and Best Practices

Edited by Carol Alexander and Elizabeth Sheedy

Introduced by David R. Koenig

The Official Handbook for the PRM Certification



Contents

[Author Biographies](#)

Introduction *David R. Koenig*

SECTION I - FINANCE THEORY, FINANCIAL INSTRUMENTS AND MARKETS

Preface I *Zvi Wiener*

A – FINANCE THEORY

I.A.1 Risk and Risk Aversion

Jacques Pezzer

I.A.1.1 Introduction

I.A.1.2 Mathematical Expectations: Prices or Utilities?

I.A.1.3 The Axiom of Independence of Choice

I.A.1.4 Maximising Expected Utility

I.A.1.4.1 The Four Basic Axioms

I.A.1.4.2 Introducing the Utility Function

I.A.1.4.3 Risk Aversion (and Risk Tolerance)

I.A.1.4.4 Certain Equivalence

I.A.1.4.5 Summary

I.A.1.5 Encoding a Utility Function

I.A.1.5.1 For an Individual

I.A.1.5.2 For a Firm

I.A.1.5.3 Ironing out Anomalies

I.A.1.6 The Mean–Variance Criterion

I.A.1.6.1 The Criterion

I.A.1.6.2 Estimating Risk Tolerance

I.A.1.6.3 Applications of the Mean–Variance Criterion

I.A.1.7 Risk-Adjusted Performance Measures

I.A.1.7.1 The Sharpe Ratio

I.A.1.7.2 RAPMs in an Equilibrium Market

I.A.1.7.2.1 The Treynor Ratio and Jensen’s Alpha

I.A.1.7.2.2 Application of the Treynor Ratio

I.A.1.7.2.3 Application of Jensen’s Alpha

I.A.1.7.3 Generalising Sharpe Ratios

I.A.1.7.3.1 The Generalised Sharpe Ratio

I.A.1.7.3.2 The Adjusted Sharpe Ratio

I.A.1.7.4 Downside RAPMs

I.A.1.7.4.1 RAROC

I.A.1.7.4.2 Sortino Ratio, Omega Index and other Kappa indices

I.A.1.8 Summary

Appendix I.A.1.A: Terminology

Appendix I.A.1.B: Utility Functions

I.A.1.B.1 The Exponential Utility Function

I.A.1.B.2 The Logarithmic Utility Function

I.A.1.B.3 The Quadratic Utility Function

I.A.1.B.4 The Power Utility Function

I.A.2 Portfolio Mathematics

Paul Glasserman

I.A.2.1 Means and Variances of Past Returns

- I.A.2.1.1 Returns
- I.A.2.1.2 Mean, Variance and Standard Deviation
- I.A.2.1.3 Portfolio Mean, Variance and Standard Deviation
- I.A.2.1.4 Correlation
- I.A.2.1.5 Correlation and Portfolio Variance
- I.A.2.1.6 Portfolio Standard Deviation
- I.A.2.2 Mean and Variance of Future Returns
 - I.A.2.2.1 Single Asset
 - I.A.2.2.2 Covariance and Correlation
 - I.A.2.2.3 Mean and Variance of a Linear Combination
 - I.A.2.2.4 Example: Portfolio Return
 - I.A.2.2.5 Example: Portfolio Profit
 - I.A.2.2.6 Example: Long and Short Positions
 - I.A.2.2.7 Example: Correlation
- I.A.2.3 Mean-Variance Tradeoffs
 - I.A.2.3.1 Achievable Expected Returns
 - I.A.2.3.2 Achievable Variance and Standard Deviation
 - I.A.2.3.3 Achievable Combinations of Mean and Standard Deviation
 - I.A.2.3.4 Efficient Frontier
 - I.A.2.3.5 Utility Maximization
 - I.A.2.3.6 Varying the Correlation Parameter
- I.A.2.4 Multiple Assets
 - I.A.2.4.1 Portfolio Mean and Variance
 - I.A.2.4.2 Vector Matrix Notation
 - I.A.2.4.3 Efficient Frontier
- I.A.2.5 A Hedging Example
 - I.A.2.5.1 Problem Formulation
 - I.A.2.5.2 Gallon-for-Gallon Hedge
 - I.A.2.5.3 Minimum-Variance Hedge
 - I.A.2.5.4 Effectiveness of the Optimal Hedge
 - I.A.2.5.5 Connection with Regression
- I.A.2.6 Serial Correlation
- I.A.2.7 Normally Distributed Returns
 - I.A.2.7.1 The Distribution of Portfolio Returns
 - I.A.2.7.2 Value-at-Risk
 - I.A.2.7.3 Probability of Reaching a Target
 - I.A.2.7.4 Probability of Beating a Benchmark

I.A.3 Capital Allocation

Keith Cuthbertson, Dirk Nitzsche

- I.A.3.1 An Overview
 - I.A.3.1.1 Portfolio Diversification
 - I.A.3.1.2 Tastes and Preferences for Risk versus Return
- I.A.3.2 Mean–Variance Criterion
- I.A.3.3 Efficient Frontier: Two Risky Assets
 - I.A.3.3.1 Different Values of the Correlation Coefficient
- I.A.3.4 Asset Allocation
 - I.A.3.4.1 The efficient frontier: n risky assets
- I.A.3.5 Combining the Risk-Free Asset with Risky Assets
- I.A.3.6 The Market Portfolio and the CML
- I.A.3.7 The Market Price of Risk and the Sharpe Ratio
- I.A.3.8 Separation Principle
- I.A.3.9 Summary
- Appendix: Mathematics of the Mean–Variance Model

I.A.4 The CAPM and Multifactor Models

Keith Cuthbertson, Dirk Nitzsche

- I.A.4.1 Overview
- I.A.4.2 Capital Asset Pricing Model
 - I.A.4.2.1 Estimating Beta
 - I.A.4.2.2 Beta and Systematic Risk
- I.A.4.3 Security Market Line
- I.A.4.4 Performance Measures
 - I.A.4.4.1 Sharpe Ratio
 - I.A.4.4.2 Jensen's 'alpha'
- I.A.4.5 The Single-Index Model
- I.A.4.6 Multifactor Models and the APT
 - I.A.4.6.1 Portfolio Returns
- I.A.4.7 Summary

I.A.5 Basics of Capital Structure

Steven Bishop

- I.A.5.1 Introduction
- I.A.5.2 Maximising Shareholder Value, Incentives and Agency Costs
 - I.A.5.2.1 Agency Costs
 - I.A.5.2.1.1 Agency Cost of Equity
 - I.A.5.2.1.2 Agency Costs of Debt
 - I.A.5.2.2 Information Asymmetries
- I.A.5.3 Characteristics of Debt and Equity
- I.A.5.4 Choice of Capital Structure
 - I.A.5.4.1 Do not think debt is attractive because the interest rate is lower than the cost of equity!
 - I.A.5.4.2 Debt can be attractive
 - I.A.5.4.2.1 Differential treatment of payments to debt-holders and shareholders
 - I.A.5.4.2.2 Greater Flexibility
 - I.A.5.4.2.3 Monitoring 'improves' performance and reduces the negative aspect of information asymmetry
 - I.A.5.4.2.4 Debt enforces a discipline of paying out operating earnings
 - I.A.5.4.2.5 Debt financing avoids negative signals about management's view of the value of equity
 - I.A.5.4.3 Debt can also be unattractive
 - I.A.5.4.3.1 Exposure to bankruptcy costs
 - I.A.5.4.3.2 Exposure to financial distress costs
 - I.A.5.4.3.3 Agency costs
 - I.A.5.4.4 Thus choose the point where disadvantages offset advantages
- I.A.5.5 Making the capital structure decision
 - I.A.5.5.1 Guidelines
 - I.A.5.5.2 What do CFOs say they consider when making a capital structure choice?
- I.A.5.6 Conclusion

I.A.6 The Term Structure of Interest Rates

Deborah Cernauskas, Elias Demetriades

- I.A.6.1 Compounding Methods
 - I.A.6.1.1 Continuous versus Discrete Compounding
 - I.A.6.1.2 Annual Compounding versus More Regular Compounding
 - I.A.6.1.3 Periodic Interest Rates versus Effective Annual Yield
- I.A.6.2 Term Structure – A Definition

- I.A.6.3 Shapes of the Yield Curve
- I.A.6.4 Spot and Forward Rates
- I.A.6.5 Term Structure Theories
 - I.A.6.5.1 Pure or Unbiased Expectations
 - I.A.6.5.2 Liquidity Preference
 - I.A.6.5.3 Market Segmentation
- I.A.6.6 Summary

I.A.7 Valuing Forward Contracts

Don Chance

- I.A.7.1 The Difference between Pricing and Valuation for Forward Contracts
- I.A.7.2 Principles of Pricing and Valuation for Forward Contracts on Assets
 - I.A.7.2.1 The Value at Time 0 of a Forward Contract
 - I.A.7.2.2 The Value at Expiration of a Forward Contract on an Asset
 - I.A.7.2.3 The Value Prior to Expiration of a Forward Contract on an Asset
 - I.A.7.2.4 The Value of a Forward Contract on an Asset when there are Cash Flows on the Asset during the Life of the Contract
 - I.A.7.2.5 Establishing the Price of a Forward Contract on an Asset
 - I.A.7.2.6 Pricing and Valuation when the Cash Flows or Holding Costs are Continuous
 - I.A.7.2.7 Numerical Examples
- I.A.7.3 Principles of Pricing and Valuation for Forward Contracts on Interest Rates
 - I.A.7.3.1 The Value of an FRA at Expiration
 - I.A.7.3.2 The Value of an FRA at the Start
 - I.A.7.3.3 The Value of an FRA During Its Life
 - I.A.7.3.4 Pricing the FRA on Day 0
 - I.A.7.3.5 Numerical Examples
- I.A.7.4 The Relationship Between Forward and Futures Prices

I.A.8 Basic Principles of Option Pricing

Paul Wilmott

- I.A.8.1 Factors Affecting Option Prices
- I.A.8.2 Put–Call Parity
- I.A.8.3 One-step Binomial Model and the Riskless Portfolio
- I.A.8.4 Delta Neutrality and Simple Delta Hedging
- I.A.8.5 Risk-Neutral Valuation
- I.A.8.6 Real versus Risk-Neutral
- I.A.8.7 The Black–Scholes–Merton Pricing Formula
- I.A.8.8 The Greeks
- I.A.8.9 Implied Volatility
- I.A.8.10 Intrinsic versus Time Value

B – FINANCIAL INSTRUMENTS

I.B.1 General Characteristics of Bonds

Lionel Martellini, Philippe Priaulet

- I.B.1.1 Definition of a Bullet Bond
- I.B.1.2 Terminology and Convention
- I.B.1.3 Market Quotes
 - I.B.1.3.1 Bond Quoted Price
 - I.B.1.3.2 Bond Quoted Yield
 - I.B.1.3.3 Bond Quoted Spread
 - I.B.1.3.4 Liquidity Spreads

- I.B.1.3.5 The Bid–Ask Spread
- I.B.1.4 Non-bullet Bonds
 - I.B.1.4.1 Strips
 - I.B.1.4.2 Floating-Rate Notes
 - I.B.1.4.3 Inflation-Indexed Bonds
- I.B.1.5 Summary

I.B.2 The Analysis of Bonds

Moorad Choudhry

- I.B.2.1 Features of Bonds
 - I.B.2.1.1 Type of Issuer
 - I.B.2.1.2 Term to Maturity
 - I.B.2.1.3 Principal and Coupon Rate
 - I.B.2.1.4 Currency
- I.B.2.2 Non-conventional Bonds
 - I.B.2.2.1 Floating-Rate Notes
 - I.B.2.2.2 Index-Linked Bonds
 - I.B.2.2.3 Zero-Coupon Bonds
 - I.B.2.2.4 Securitised Bonds
 - I.B.2.2.5 Bonds with Embedded Options
- I.B.2.3 Pricing a Conventional Bond
 - I.B.2.3.1 Bond Cash Flows
 - I.B.2.3.2 The Discount Rate
 - I.B.2.3.3 Conventional Bond Pricing
 - I.B.2.3.4 Pricing Undated Bonds
 - I.B.2.3.5 Pricing Conventions
 - I.B.2.3.6 Clean and Dirty Bond Prices: Accrued Interest
- I.B.2.4 Market Yield
 - I.B.2.4.1 Yield Measurement
 - I.B.2.4.2 Current Yield
 - I.B.2.4.3 Yield to Maturity
- I.B.2.5 Relationship between Bond Yield and Bond Price
- I.B.2.6 Duration
 - I.B.2.6.1 Calculating Macaulay Duration and Modified Duration
 - I.B.2.6.2 Properties of the Macaulay Duration
 - I.B.2.6.3 Properties of the Modified Duration
- I.B.2.7 Hedging Bond Positions
- I.B.2.8 Convexity
- I.B.2.9 A Summary of Risks Associated with Bonds

I.B.3 Futures and Forwards

Keith Cuthbertson, Dirk Nitzsche

- I.B.3.1 Introduction
- I.B.3.2 Stock Index Futures
 - I.B.3.2.1 Contract Specifications
 - I.B.3.2.2 Index arbitrage and program trading
 - I.B.3.2.3 Hedging Using Stock Index Futures
 - I.B.3.2.4 Tailing the Hedge
 - I.B.3.2.5 Summary
- I.B.3.3 Currency Forwards and Futures
 - I.B.3.3.1 Currency Forward Contracts
 - I.B.3.3.2 Currency Futures Contracts
 - I.B.3.3.3 Hedging Currency Futures and Forwards
 - I.B.3.3.4 Summary

- I.B.3.4 Commodity Futures
- I.B.3.5 Forward Rate Agreements
 - I.B.3.5.1 Settlement Procedures
- I.B.3.6 Short-Term Interest-Rate Futures
 - I.B.3.6.1 US T-bill Futures
 - I.B.3.6.2 Three-Month Eurodollar Futures
 - I.B.3.6.3 Sterling Three-Month Futures
 - I.B.3.6.4 Hedging Interest-Rate Futures
 - I.B.3.6.5 Hedge Ratios
 - I.B.3.6.6 Hedging Using US T-bill Futures
 - I.B.3.6.7 Summary
- I.B.3.7 T-bond Futures
 - I.B.3.7.1 Contract Specifications
 - I.B.3.7.1.1 UK Long Gilt Futures Contract
 - I.B.3.7.1.2 US T-bond Futures Contract
 - I.B.3.7.2 Conversion Factor and Cheapest to Deliver
 - I.B.3.7.3 Hedging Using T-bond Futures
 - I.B.3.7.4 Hedging a Single Bond
 - I.B.3.7.5 Hedging a Portfolio of Bonds
 - I.B.3.7.6 Summary
- I.B.3.8 Stack and Strip Hedges
- I.B.3.9 Concluding Remarks

I.B.4 Swaps

Salih Neftci

- I.B.4.1 What is a Swap?
- I.B.4.2 Types of Swaps
 - I.B.4.2.1 Equity Swaps
 - I.B.4.2.2 Commodity Swaps
 - I.B.4.2.3 Interest Rate Swaps
 - I.B.4.2.4 Currency Swaps
 - I.B.4.2.5 Basis Swaps
 - I.B.4.2.6 Volatility Swaps
- I.B.4.3 Engineering Interest-Rate Swaps
- I.B.4.4 Risk of Swaps
 - I.B.4.4.1 Market Risk
 - I.B.4.4.2 Credit Risk and Counterparty Risk
 - I.B.4.4.3 Volatility and Correlation Risk
- I.B.4.5 Other Swaps
- I.B.4.6 Uses of Swaps
 - I.B.4.6.1 Uses of Equity Swaps
- I.B.4.7 Swap Conventions

I.B.5 Vanilla Options

Paul Wilmott

- I.B.5.1 Stock Options – Characteristics and Payoff Diagrams
- I.B.5.2 American versus European Options
- I.B.5.3 Strategies Involving a Single Option and a Stock
- I.B.5.4 Spread Strategies
 - I.B.5.4.1 Bull and Bear Spreads
 - I.B.5.4.2 Calendar Spreads
- I.B.5.5 Other Strategies
 - I.B.5.5.1 Straddles and Strangles
 - I.B.5.5.2 Risk Reversal

- I.B.5.5.3 Collars
- I.B.5.5.4 Butterflies and Condors

I.B.6 Credit Derivatives

Moorad Choudhry

- I.B.6.1 Introduction
 - I.B.6.1.1 Why Use Credit Derivatives?
 - I.B.6.1.2 Classification of Credit Derivative Instruments
 - I.B.6.1.3 Definition of a Credit Event
- I.B.6.2 Credit Default Swaps
- I.B.6.3 Credit-Linked Notes
- I.B.6.4 Total Return Swaps
 - I.B.6.4.1 Synthetic Repo
 - I.B.6.4.2 Reduction in Credit Risk
 - I.B.6.4.3 Capital Structure Arbitrage
 - I.B.6.4.4 The TRS as a Funding Instrument
- I.B.6.5 Credit Options
- I.B.6.6 Synthetic Collateralised Debt Obligations
 - I.B.6.6.1 Cash Flow CDOs
 - I.B.6.6.2 What is a Synthetic CDO?
 - I.B.6.6.3 Funding Synthetic CDOs
 - I.B.6.6.4 Variations in Synthetic CDOs
 - I.B.6.6.5 Use of Synthetic CDOs
 - I.B.6.6.6 Advantages and Limitations of Synthetic Structures
- I.B.6.7 General Applications of Credit Derivatives
 - I.B.6.7.1 Use of Credit Derivatives by Portfolio Managers
 - I.B.6.7.1.1 Enhancing portfolio returns
 - I.B.6.7.1.2 Reducing credit exposure
 - I.B.6.7.1.3 Credit switches and zero-cost credit exposure
 - I.B.6.7.1.4 Exposure to market sectors
 - I.B.6.7.1.5 Trading Credit spreads
 - I.B.6.7.2 Use of Credit Derivatives by Banks
- I.B.6.8 Unintended Risks in Credit Derivatives
- I.B.6.9 Summary

I.B.7 Caps, Floors & Swaptions

Lionel Martellini, Philippe Priaulet

- I.B.7.1 Caps, Floors and Collars: Definition and Terminology
- I.B.7.2 Pricing Caps, Floors and Collars
 - I.B.7.2.1 Cap Formula
 - I.B.7.2.2 Floor Formula
 - I.B.7.2.3 Market Quotes
- I.B.7.3 Uses of Caps, Floors and Collars
 - I.B.7.3.1 Limiting the Financial Cost of Floating-Rate Liabilities
 - I.B.7.3.2 Protecting the Rate of Return of a Floating-Rate Asset
- I.B.7.4 Swaptions: Definition and Terminology
- I.B.7.5 Pricing Swaptions
 - I.B.7.5.1 European Swaption Pricing Formula
 - I.B.7.5.2 Market Quotes
- I.B.7.6 Uses of Swaptions
- I.B.7.7 Summary

I.B.8 Convertible Bonds

Izzy Nelken

- I.B.8.1 Introduction
 - I.B.8.1.1 Convertibles – a definition
 - I.B.8.1.2 Convertible Bond Market Size
 - I.B.8.1.3 A Brief History
- I.B.8.2 Characteristics of Convertibles
 - I.B.8.2.1 Relationship with Stock Price
 - I.B.8.2.2 Call and Put Features
 - I.B.8.2.3 Players in the Convertible Bond Market
 - I.B.8.2.4 Convertible Bond Funds
 - I.B.8.2.5 Convertible Arbitrage Hedge Funds
- I.B.8.3 Capital Structure Implications (for Banks)
- I.B.8.4 Mandatory Convertibles
- I.B.8.5 Valuation and Risk Assessment
- I.B.8.6 Summary

I.B.9 Simple Exotics

Catriona March

- I.B.9.1 Introduction
- I.B.9.2 A Short History
- I.B.9.3 Classifying Exotics
- I.B.9.4 Notation
- I.B.9.5 Digital Options
 - I.B.9.5.1 Cash-or-Nothing Options
 - I.B.9.5.2 Asset-or-Nothing Options
 - I.B.9.5.3 Vanillas and Digitals as Building Blocks
 - I.B.9.5.4 Contingent Premium Options
 - I.B.9.5.5 Range Notes
 - I.B.9.5.6 Managing Digital Options
- I.B.9.6 Two Asset Options
 - I.B.9.6.1 Product and Quotient Options
 - I.B.9.6.2 Exchange Options
 - I.B.9.6.3 Outperformance Options
 - I.B.9.6.4 Other Two-Colour Rainbow Options
 - I.B.9.6.5 Spread Options
 - I.B.9.6.6 Correlation Risk
- I.B.9.7 Quantos
 - I.B.9.7.1 Foreign Asset Option Struck in Foreign Currency
 - I.B.9.7.2 Foreign Asset Option Struck in Domestic Currency
 - I.B.9.7.3 Implied Correlation
 - I.B.9.7.4 Foreign Asset Linked Currency Option
 - I.B.9.7.5 Guaranteed Exchange Rate Foreign Asset Options
- I.B.9.8 Second-Order Contracts
 - I.B.9.8.1 Compound Options
 - I.B.9.8.2 Typical Uses of Compound Options
 - I.B.9.8.3 Instalment Options
 - I.B.9.8.4 Extendible Options
- I.B.9.9 Decision Options
 - I.B.9.9.1 American Options
 - I.B.9.9.2 Bermudan Options
 - I.B.9.9.3 Shout Options
- I.B.9.10 Average Options
 - I.B.9.10.1 Average Rate and Average Strike Options

- I.B.9.10.2 Motivations and Uses
- I.B.9.10.3 Other Options Involving Averages
- I.B.9.10.4 Pricing and Hedging Average Options
- I.B.9.11 Options on Baskets of Assets
 - I.B.9.11.1 Basket Options
 - I.B.9.11.2 Pricing and Hedging Basket Options
 - I.B.9.11.3 Mountain Options
- I.B.9.12 Barrier and Related Options
 - I.B.9.12.1 Single-Barrier Options
 - I.B.9.12.2 No-Touch, One-Touch and Rebates
 - I.B.9.12.3 Partial-Barrier Options
 - I.B.9.12.4 Double-Barrier Options
 - I.B.9.12.5 Even More Barrier Options
 - I.B.9.12.6 Relationships
 - I.B.9.12.7 Ladders
 - I.B.9.12.8 Lookback and Hindsight Options
- I.B.9.13 Other Path-Dependent Options
 - I.B.9.13.1 Forward Start Options
 - I.B.9.13.2 Reset Options
 - I.B.9.13.3 Cliquet Options
- I.B.9.14 Resolution Methods
- I.B.9.15 Summary

C - MARKETS

I.C.1 The Structure of Financial Markets

Colin Lawrence, Alistair Milne

- I.C.1.1 Introduction
- I.C.1.2 Global Markets and Their Terminology
- I.C.1.3 Drivers of Liquidity
 - I.C.1.3.1 Repo Markets
- I.C.1.4 Liquidity and Financial Risk Management
- I.C.1.5 Exchanges versus OTC Markets
- I.C.1.6 Technological Change
- I.C.1.7 Post-trade Processing
- I.C.1.8 Retail and Wholesale Brokerage
- I.C.1.9 New Financial Markets
- I.C.1.10 Conclusion

I.C.2 The Money Markets

Canadian Securities Institute

- I.C.2.1 Introduction
- I.C.2.2 Characteristics of Money Market Instruments
- I.C.2.3 Deposits and Loans
 - I.C.2.3.1 Deposits from Businesses
 - I.C.2.3.2 Loans to Businesses
 - I.C.2.3.3 Repurchase Agreements
 - I.C.2.3.4 International Markets
 - I.C.2.3.5 The London Interbank Offered Rate (LIBOR)
- I.C.2.4 Money Market Securities
 - I.C.2.4.1 Treasury Bills
 - I.C.2.4.2 Commercial Paper
 - I.C.2.4.3 Bankers' Acceptances
 - I.C.2.4.4 Certificates of Deposit

I.C.2.5 Summary

I.C.3 The Bond Market

Moorad Choudhry, Lionel Martellini, Philippe Priaulet

I.C.3.1 Introduction

I.C.3.2 The Players

I.C.3.2.1 Intermediaries and Banks

I.C.3.2.2 Institutional Investors

I.C.3.2.3 Market Professionals

I.C.3.3 Bonds by Issuers

I.C.3.3.1 Government Bonds

I.C.3.3.2 US Agency Bonds

I.C.3.3.3 Municipal Bonds

I.C.3.3.4 Corporate Bonds

I.C.3.3.5 Eurobonds (International Bonds)

I.C.3.4 The Markets

I.C.3.4.1 The Government Bond Market

I.C.3.4.2 The Corporate Bond Market

I.C.3.4.2.1 The market by country and sector

I.C.3.4.2.2 Underwriting a new issue

I.C.3.4.3 The Eurobond Market

I.C.3.4.4 Market Conventions

I.C.3.5 Credit Risk

I.C.3.6 Summary

I.C.4 The Foreign Exchange Market

Canadian Securities Institute, Toronto

I.C.4.1 Introduction

I.C.4.2 The Interbank Market

I.C.4.3 Exchange-Rate Quotations

I.C.4.3.1 Direct Dealing

I.C.4.3.2 Foreign Exchange Brokers

I.C.4.3.3 Electronic Brokering Systems

I.C.4.3.4 The Role of the US Dollar

I.C.4.3.5 Market and Quoting Conventions

I.C.4.3.6 Cross Trades and Cross Rates

I.C.4.4 Determinants of Foreign Exchange Rates

I.C.4.4.1 The Fundamental Approach

I.C.4.4.2 A Short-Term Approach

I.C.4.4.3 Central Bank Intervention

I.C.4.5 Spot and Forward Markets

I.C.4.5.1 The Spot Market

I.C.4.5.2 The Forward Market

I.C.4.5.2.1 Forward Discounts and Premiums

I.C.4.5.2.2 Interest-Rate Parity

I.C.4.6 Structure of a Foreign Exchange Operation

I.C.4.7 Summary/Conclusion

I.C.5 The Stock Market

Andrew Street

I.C.5.1 Introduction

I.C.5.2 The Characteristics of Common Stock

I.C.5.2.1 Share Premium and Capital Accounts and Limited Liability

- I.C.5.2.2 Equity Shareholder's Rights and Dividends
- I.C.5.2.3 Other Types of Equity Shares – Preference Shares
- I.C.5.2.4 Equity Price Data
- I.C.5.2.5 Market Capitalisation (or 'Market Cap')
- I.C.5.2.6 Stock Market Indices
- I.C.5.2.7 Equity Valuation
- I.C.5.3 Stock Markets and their Participants
 - I.C.5.3.1 The Main Participants – Firms, Investment Banks and Investors
 - I.C.5.3.2 Market Mechanics
- I.C.5.4 The Primary Market – IPOs and Private Placements
 - I.C.5.4.1 Basic Primary Market Process
 - I.C.5.4.2 Initial Public Offerings
 - I.C.5.4.3 Private Placements
- I.C.5.5 The Secondary Market – the Exchange versus OTC Market
 - I.C.5.5.1 The Exchange
 - I.C.5.5.2 The Over-the-Counter Market
 - I.C.5.6 Trading Costs
 - I.C.5.6.1 Commissions
 - I.C.5.6.2 Bid–Offer Spread
 - I.C.5.6.3 Market Impact
- I.C.5.7 Buying on Margin
 - I.C.5.7.1 Leverage
 - I.C.5.7.2 Percentage Margin and Maintenance Margin
 - I.C.5.7.3 Why Trade on Margin?
- I.C.5.8 Short Sales and Stock Borrowing Costs
 - I.C.5.8.1 Short Sale
 - I.C.5.8.2 Stock Borrowing
- I.C.5.9 Exchange-Traded Derivatives on Stocks
 - I.C.5.9.1 Single Stock and Index Options
 - I.C.5.9.2 Expiration Dates
 - I.C.5.9.3 Strike Prices
 - I.C.5.9.4 Flex Options
 - I.C.5.9.5 Dividends and Corporate Actions
 - I.C.5.9.6 Position Limits
 - I.C.5.9.7 Trading
 - I.C.5.10 Summary

I.C.6 The Futures Market

Canadian Securities Institute

- I.C.6.1 Introduction
- I.C.6.2 History of Forward-Based Derivatives and Futures Markets
- I.C.6.3 Futures Contracts and Markets
 - I.C.6.3.1 General Characteristics of Futures Contracts and Markets
 - I.C.6.3.2 Settlement of Futures Contracts
 - I.C.6.3.3 Types of Orders
 - I.C.6.3.4 Margin Requirements and Marking to Market
 - I.C.6.3.5 Leverage
 - I.C.6.3.6 Reading a Futures Quotation Page
 - I.C.6.3.7 Liquidity and Trading Costs
- I.C.6.4 Options on Futures
- I.C.6.5 Futures Exchanges and Clearing Houses
 - I.C.6.5.1 Exchanges
 - I.C.6.5.2 Futures Exchange Functions
 - I.C.6.5.3 Clearing Houses
 - I.C.6.5.4 Marking-to-Market and Margin

- I.C.6.6 Market Participants – Hedgers
- I.C.6.7 Market Participants – Speculators
 - I.C.6.7.1 Locals
 - I.C.6.7.2 Day Traders
 - I.C.6.7.3 Position Traders
 - I.C.6.7.4 Spreaders
 - I.C.6.7.4.1 Intramarket Spreads
 - I.C.6.7.4.2 Intercommodity Spreads
 - I.C.6.7.4.3 Intermarket Spreads
 - I.C.6.7.4.4 Commodity Product Spread
- I.C.6.8 Market Participants – Managed Futures Investors
- I.C.6.9 Summary and Conclusion

I.C.7 The Structure of Commodities Markets

Colin Lawrence, Alistair Milne

- I.C.7.1 Introduction
- I.C.7.2 The Commodity Universe and Anatomy of Markets
 - I.C.7.2.1 Commodity Types and Characteristics
 - I.C.7.2.2 The Markets for Trading
 - I.C.7.2.3 Delivery and Settlement Methods
 - I.C.7.2.4 Commodity Market Liquidity
 - I.C.7.2.5 The Special Case of Gold as a Reserve Asset
- I.C.7.3 Spot–Forward Pricing Relationships
 - I.C.7.3.1 Backwardation and Contango
 - I.C.7.3.2 Reasons for Backwardation
 - I.C.7.3.3 The No-Arbitrage Condition
- I.C.7.4 Short Squeezes, Corners and Regulation
 - I.C.7.4.1 Historical Experience
 - I.C.7.4.2 The Exchange Limits
- I.C.7.5 Risk Management at the Commodity Trading Desk
- I.C.7.6 The Distribution of Commodity Returns
 - I.C.7.6.1 Evidence of Non-normality
 - I.C.7.6.2 What Drives Commodity Prices?
- I.C.7.7 Conclusions

I.C.8 The Energy Markets

Peter Fusaro

- I.C.8.1 Introduction
- I.C.8.2 Market Overview
 - I.C.8.2.1 The Products
 - I.C.8.2.2 The Risks
 - I.C.8.2.3 Developing a Cash Market
- I.C.8.3 Energy Futures Markets
 - I.C.8.3.1 The Exchanges
 - I.C.8.3.2 The Contracts
 - I.C.8.3.3 Options on Energy Futures
 - I.C.8.3.4 Hedging in Energy Futures Markets
 - I.C.8.3.5 Physical Delivery
 - I.C.8.3.6 Market Changes: Backwardation and Contango
- I.C.8.4 OTC Energy Derivative Markets
 - I.C.8.4.1 The Singapore Market
 - I.C.8.4.2 The European Energy Markets
 - I.C.8.4.3 The North American Markets
- I.C.8.5 Emerging Energy Commodity Markets

- I.C.8.5.1 Coal Trading
- I.C.8.5.2 Weather Derivatives
- I.C.8.5.3 Green Trading
- I.C.8.5.4 Freight-Rate Swaps
- I.C.8.5.5 Derivative Forward Price Curves
- I.C.8.6 The Future of Energy Trading
 - I.C.8.6.1 Re-emergence of Speculative Trading?
 - I.C.8.6.2 Electronic Energy Trading
 - I.C.8.6.3 Trading in Asian Markets
- I.C.8.7 Conclusion

SECTION II – MATHEMATICAL FOUNDATIONS OF RISK MEASUREMENT

Preface II *Carol Alexander*

II.A Foundations

Keith Parramore, Terry Watsham

- II.A.1 Symbols and Rules
 - II.A.1.1 Expressions, Functions, Graphs, Equations and Greek
 - II.A.1.2 The Algebra of Number
 - II.A.1.3 The Order of Operations
- II.A.2 Sequences and Series
 - II.A.2.1 Sequences
 - II.A.2.2 Series
- II.A.3 Exponents and Logarithms
 - II.A.3.1 Exponents
 - II.A.3.2 Logarithms
 - II.A.3.3 The Exponential Function and Natural Logarithms
- II.A.4 Equations and Inequalities
 - II.A.4.1 Linear Equations in One Unknown
 - II.A.4.2 Inequalities
 - II.A.4.3 Systems of Linear Equations in More Than One Unknown
 - II.A.4.4 Quadratic Equations
- II.A.5 Functions and Graphs
 - II.A.5.1 Functions
 - II.A.5.2 Graphs
 - II.A.5.3 The Graphs of Some Functions
- II.A.6 Case Study – Continuous Compounding
 - II.A.6.1 Repeated Compounding
 - II.A.6.2 Discrete versus Continuous Compounding
- II.A.7 Summary

II.B Descriptive Statistics

Keith Parramore, Terry Watsham

- II.B.1 Introduction
- II.B.2 Data
 - II.B.2.1 Continuous and Discrete Data
 - II.B.2.2 Grouped Data
 - II.B.2.3 Graphical Representation of Data
 - II.B.2.3.1 The Frequency Bar Chart
 - II.B.2.3.2 The Relative Frequency Distribution
 - II.B.2.3.3 The Cumulative Frequency Distribution
 - II.B.2.3.4 The Histogram
- II.B.3 The Moments of a Distribution

- II.B.4 Measures of Location or Central Tendency – Averages
 - II.B.4.1 The Arithmetic Mean
 - II.B.4.2 The Geometric Mean
 - II.B.4.3 The Median and the Mode
- II.B.5 Measures of Dispersion
 - II.B.5.1 Variance
 - II.B.5.2 Standard Deviation
 - II.B.5.3 Case Study: Calculating Historical Volatility from Returns Data
 - II.B.5.4 The Negative Semi-variance and Negative Semi-deviation
 - II.B.5.5 Skewness
 - II.B.5.6 Kurtosis
- II.B.6 Bivariate Data
 - II.B.6.1 Covariance
 - II.B.6.2 The Covariance Matrix
 - II.B.6.3 The Correlation Coefficient
 - II.B.6.4 The Correlation Matrix
 - II.B.6.5 Case Study: Calculating the Volatility of a Portfolio

II.C Calculus

Keith Parramore, Terry Watsham

- II.C.1 Differential Calculus
 - II.C.1.1 Functions
 - II.C.1.2 The First Derivative
 - II.C.1.3 Notation
 - II.C.1.4 Simple Rules
 - II.C.1.4.1 Differentiating Constants
 - II.C.1.4.2 Differentiating a Linear Function
 - II.C.1.4.3 The Gradient of a Straight Line
 - II.C.1.4.4 The Derivative of a Power of x
 - II.C.1.4.5 Differentiating a scalar multiple of a function
 - II.C.1.4.6 Differentiating the Sum of Two Functions of x
 - II.C.1.4.7 Differentiating the Product of Two Functions of x
 - II.C.1.4.8 Differentiating the Quotient of Two Functions of x
 - II.C.1.4.9 Differentiating a Function of a Function
 - II.C.1.4.10 Differentiating the Exponential Function
 - II.C.1.4.11 Differentiating the Natural Logarithmic Function
- II.C.2 Case Study: Modified Duration of a Bond
- II.C.3 Higher-Order Derivatives
 - II.C.3.1 Second Derivatives
 - II.C.3.2 Further Derivatives
 - II.C.3.3 Taylor Approximations
- II.C.4 Financial Applications of Second Derivatives
 - II.C.4.1 Convexity
 - II.C.4.2 Convexity in Action
 - II.C.4.3 The Delta and Gamma of an Option
- II.C.5 Differentiating a Function of More than One Variable
 - II.C.5.1 Partial Differentiation
 - II.C.5.2 Total differentiation
- II.C.6 Integral Calculus
 - II.C.6.1 Indefinite and Definite Integrals
 - II.C.6.2 Rules for Integration
 - II.C.6.3 Guessing
- II.C.7 Optimisation
 - II.C.7.1 Finding the Minimum or Maximum of a Function of One Variable
 - II.C.7.2 Maxima and Minima of Functions of More than One Variable

- II.C.7.3 Optimization Subject to Constraints: Lagrange Multipliers
- II.C.7.4 Applications

II.D Linear Algebra

Keith Parramore, Terry Watsham

- II.D.1 Matrix Algebra
 - II.D.1.1 Matrices
 - II.D.1.2 Vectors and Transposes
 - II.D.1.3 Manipulation of Matrices
 - II.D.1.4 Matrix Multiplication
 - II.D.1.5 Inverting a Matrix
- II.D.2 Application of Matrix Algebra to Portfolio Construction
 - II.D.2.1 Calculating the Risk of an Existing Portfolio
 - II.D.2.2 Deriving Asset Weights for the Minimum Risk Portfolio
 - II.D.2.3 Hedging a Vanilla Option Position
 - II.D.2.3.1 Calculating the position delta
 - II.D.2.3.2 Establishing the delta-neutral hedge
 - II.D.2.3.3 Gamma neutrality
 - II.D.2.3.4 Vega neutrality
 - II.D.2.3.5 Hedging a short option position
- II.D.3 Quadratic Forms
 - II.D.3.1 The Variance of Portfolio Returns as a Quadratic Form
 - II.D.3.2 Definition of Positive Definiteness
- II.D.4 Cholesky Decomposition
 - II.D.4.1 The Cholesky Arithmetic
 - II.D.4.2 Simulation in Excel
- II.D.5 Eigenvalues and Eigenvectors
 - II.D.5.1 Matrices as Transformations
 - II.D.5.2 Definition of Eigenvector and Eigenvalue
 - II.D.5.3 Determinants
 - II.D.5.4 The Characteristic Equation
 - II.D.5.4.1 Testing for Positive Semi-definiteness
 - II.D.5.4.2 Using the characteristic equation to find the eigenvalues of a covariance matrix
 - II.D.5.4.3 Eigenvalues and eigenvectors of covariance and correlation matrices
 - II.D.5.5 Principal Components

II.E Probability Theory in Finance

Keith Parramore, Terry Watsham

- II.E.1 Definitions and Rules
 - II.E.1.1 Definitions
 - II.E.1.1.1 The classical approach
 - II.E.1.1.2 The Bayesian approach
 - II.E.1.2 Rules for Probability
 - II.E.1.2.1 (A or B) and (A and B)
 - II.E.1.2.2 Conditional Probability
- II.E.2 Probability Distributions
 - II.E.2.1 Random Variables
 - II.E.2.1.1 Discrete Random Variables
 - II.E.2.1.2 Continuous Random Variables
 - II.E.2.2 Probability Density Functions and Histograms
 - II.E.2.3 The Cumulative Distribution Function
 - II.E.2.4 The Algebra of Random Variables
 - II.E.2.4.1 Scalar Multiplication of a Random Variable

- II.E.2.5 The Expected Value of a Discrete Random Variable
- II.E.2.6 The Variance of a Discrete Random Variable
- II.E.2.7 The Algebra of Continuous Random Variables
- II.E.3 Joint Distributions
 - II.E.3.1 Bivariate Random Variables
 - II.E.3.2 Covariance
 - II.E.3.3 Correlation
 - II.E.3.4 The Expected Value and Variance of a Linear Combination of Random Variables
- II.E.4 Specific Probability Distributions
 - II.E.4.1 The Binomial Distribution
 - II.E.4.1.1 Calculating the ‘Number of Ways’
 - II.E.4.1.2 Calculating the Probability of r Successes
 - II.E.4.1.3 Expectation and Variance
 - II.E.4.2 The Poisson Distribution
 - II.E.4.2.1 Illustrations
 - II.E.4.2.2 Expectation and Variance
 - II.E.4.3 The Uniform Continuous Distribution
 - II.E.4.4.1 Normal Curves
 - II.E.4.4.2 The Standard Normal Probability Density Function
 - II.E.4.4.3 Finding Areas under a Normal Curve Using Excel
 - II.E.4.5 The Lognormal Probability Distribution
 - II.E.4.5.1 Lognormal Curves
 - II.E.4.5.2 The Lognormal Distribution Applied to Asset Prices
 - II.E.4.5.3 The Mean and Variance of the Lognormal Distribution
 - II.E.4.5.4 Application of the Lognormal Distribution to Future Asset Prices [*not in PRM exam*]
 - II.E.4.6 Student’s t Distribution
 - II.E.4.7 The Bivariate Normal Distribution

II.F Regression

Keith Parramore, Terry Watsham

- II.F.1 Simple Linear Regression
 - II.F.1.1 The Model
 - II.F.1.2 The Scatter Plot
 - II.F.1.3 Estimating the Parameters
- II.F.2 Multiple Linear Regression
 - II.F.2.1 The model
 - II.F.2.2 Estimating the Parameters
- II.F.3 Evaluating the Regression Model
 - II.F.3.1 Intuitive Interpretation
 - II.F.3.2 Adjusted R²
 - II.F.3.3 Testing for Statistical Significance
- II.F.4 Confidence Intervals
 - II.F.4.1 Confidence Intervals for the Regression Parameters
- II.F.5 Hypothesis Testing
 - II.F.5.1 Significance Tests for the Regression Parameters
 - II.F.5.2 Significance Test for R²
 - II.F.5.3 Type I and type II errors
- II.F.6 Prediction
- II.F.7 Breakdown of the OLS Assumptions
 - II.F.7.1 Heteroscedasticity
 - II.F.7.2 Autocorrelation
 - II.F.7.3 Multicollinearity
- II.F.8 Random Walks and Mean-Reversion

- II.F.9 Maximum Likelihood Estimation
- II.F.10 Summary

II.G Numerical Methods

Keith Parramore, Terry Watsham

- II.G.1 Solving (Non-differential) Equations
 - II.G.1.1 Three Problems
 - II.G.1.2 Bisection
 - II.G.1.3 Newton–Raphson
 - II.G.1.4 Goal Seek
- II.G.2 Numerical Optimisation
 - II.G.2.1 The Problem
 - II.G.2.2 Unconstrained Numerical Optimisation
 - II.G.2.3 Constrained Numerical Optimisation
 - II.G.2.4 Portfolio Optimisation Revisited
- II.G.3 Numerical Methods for Valuing Options
 - II.G.3.1 Binomial Lattices
 - II.G.3.2 Finite Difference Methods
 - II.G.3.3 Simulation
- II.G.4 Summary

SECTION III – RISK MANAGEMENT PRACTICES

Preface III *Elizabeth Shedy*

III.0 Capital Allocation and Risk Adjusted Performance

Andrew Aziz, Dan Rosen

- III.0.1 Introduction
 - III.0.1.1 Role of Capital in Financial Institution
 - III.0.1.2 Types of Capital
 - III.0.1.3 Capital as a Management Tool
- III.0.2 Economic Capital
 - III.0.2.1 Understanding Economic Capital
 - III.0.2.2 The Top-Down Approach to Calculating Economic Capital
 - III.0.2.2.1 Top-Down Earnings Volatility Approach
 - III.0.2.2.2 Top-Down Option-Theoretic Approach
 - III.0.2.3 The Bottom-Up Approach to Calculating Economic Capital
 - III.0.2.4 Stress Testing of Portfolio Losses and Economic Capital
 - III.0.2.5 Enterprise Capital Practices – Aggregation
 - III.0.2.6 Economic Capital as Insurance for the Value of the Firm
- III.0.3 Regulatory Capital
 - III.0.3.1 Regulatory Capital Principles
 - III.0.3.2 The Basel Committee of Banking Supervision and the Basel Accord
 - III.0.3.3 Basel I Regulation
 - III.0.3.3.1 Minimum Capital Requirements under Basel I
 - III.0.3.3.2 Regulatory Arbitrage under Basel I
 - III.0.3.3.3 Meeting Capital Adequacy Requirements
 - III.0.3.4 Basel II Accord – Latest Proposals
 - III.0.3.4.1 Pillar 1 - Minimum Capital Requirements
 - III.0.3.4.2 Pillar 2 - Supervisory Review
 - III.0.3.4.3 Pillar 3 - Market Discipline
 - III.0.3.5 A Simple Derivation of Regulatory Capital
- III.0.4 Capital Allocation and Risk Contributions

- III.0.4.1 Capital Allocation
- III.0.4.2 Risk Contribution Methodologies for EC Allocation
 - III.0.4.2.1 Stand-alone EC Contributions
 - III.0.4.2.2 Incremental EC Contributions
 - III.0.4.2.3 Marginal EC Contributions
 - III.0.4.2.4 Alternative Methods for Additive Contributions
- III.0.5 RAROC and Risk-Adjusted Performance
 - III.0.5.1 Objectives of RAPM
 - III.0.5.2 Mechanics of RAROC
 - III.0.5.3 RAROC and Capital Allocation Methodologies
- III.0.6 Summary and Conclusions

A – MARKET RISK

III.A.1 Market Risk Management

Jacques Pezier

- III.A.1.1 Introduction
- III.A.1.2 Market Risk
 - III.A.1.2.1 Why is Market Risk Management Important?
 - III.A.1.2.2 Distinguishing Market Risk from Other Risks
- III.A.1.3 Market Risk Management Tasks
- III.A.1.4 The Organisation of Market Risk Management
- III.A.1.5 Market Risk Management in Fund Management
 - III.A.1.5.1 Market Risk in Fund Management
 - III.A.1.5.2 Identification
 - III.A.1.5.3 Assessment
 - III.A.1.5.4 Control/Mitigation
- III.A.1.6 Market Risk Management in Banking
 - III.A.1.6.1 Market Risk in Banking
 - III.A.1.6.2 Identification
 - III.A.1.6.3 Assessment
 - III.A.1.6.4 Control/Mitigation
- III.A.1.7 Market Risk Management in Non-financial Firms
 - III.A.1.7.1 Market Risk in Non-Financial Firms
 - III.A.1.7.2 Identification
 - III.A.1.7.3 Assessment
 - III.A.1.7.4 Control/Mitigation
- III.A.1.8 Summary

III.A.2 Introduction to Value at Risk Models

Kevin Dowd, David Rowe

- III.A.2.1 Introduction
- III.A.2.2 Definition of VaR
- III.A.2.3 Internal Models for Market Risk Capital
- III.A.2.4 Analytical VaR Models
- III.A.2.5 Monte Carlo Simulation VaR
 - III.A.2.5.1 Methodology
 - III.A.2.5.2 Applications of Monte Carlo simulation
 - III.A.2.5.3 Advantages and Disadvantages of Monte Carlo VaR
- III.A.2.6 Historical Simulation VaR
 - III.A.2.6.1 The Basic Method
 - III.A.2.6.2 Weighted historical simulation
 - III.A.2.6.3 Advantages and Disadvantages of Historical Approaches
- III.A.2.7 Mapping Positions to Risk Factors

- III.A.2.7.1 Mapping Spot Positions
- III.A.2.7.2 Mapping Equity Positions
- III.A.2.7.3 Mapping Zero-Coupon Bonds
- III.A.2.7.4 Mapping Forward/Futures Positions
- III.A.2.7.5 Mapping Complex Positions
- III.A.2.7.6 Mapping Options: Delta and Delta-Gamma Approaches
- III.A.2.8 Backtesting VaR Models
- III.A.2.9 Why Financial Markets Are Not 'Normal'
- III.A.2.10 Summary

III.A.3 Advanced Value at Risk Models

Carol Alexander, Elizabeth Sheedy

- III.A.3.1 Introduction
- III.A.3.2 Standard Distributional Assumptions
- III.A.3.3 Models of Volatility Clustering
 - III.A.3.3.1 Exponentially Weighted Moving Average (EWMA)
 - III.A.3.3.2 GARCH Models
- III.A.3.4 Volatility Clustering and VaR
 - III.A.3.4.1 VaR using EWMA
 - III.A.3.4.2 VaR and GARCH
- III.A.3.5 Alternative Solutions to Non-Normality
 - III.A.3.5.1 VaR with the Student's-t distribution
 - III.A.3.5.2 VaR with EVT
 - III.A.3.5.3 VaR with Normal Mixtures
- III.A.3.6 Decomposition of VaR
 - III.A.3.6.1 Stand Alone Capital
 - III.A.3.6.2 Incremental VaR
 - III.A.3.6.3 Marginal Capital
- III.A.3.7 Principal Component Analysis
 - III.A.3.7.1 PCA in Action
 - III.A.3.7.2 VaR with PCA
- III.A.3.8 Summary

III.A.4 Stress Testing

Barry Schachter

- III.A.4.1 Introduction
- III.A.4.2 Historical Context
- III.A.4.3 Conceptual Context
- III.A.4.4 Stress Testing in Practice
- III.A.4.5 Approaches to Stress Testing: An Overview
- III.A.4.6 Historical Scenarios
 - III.A.4.6.1 Choosing Event Periods
 - III.A.4.6.2 Specifying Shock Factors
 - III.A.4.6.3 Missing Shock Factors
- III.A.4.7 Hypothetical Scenarios
 - III.A.4.7.1 Modifying the Covariance Matrix
 - III.A.4.7.2 Specifying Factor Shocks (to 'create' an event)
 - III.A.4.7.3 Systemic Events and Stress-Testing Liquidity
 - III.A.4.7.4 Sensitivity Analysis
 - III.A.4.7.5 Hybrid Methods
- III.A.4.8 Algorithmic Approaches to Stress Testing
 - III.A.4.8.1 Factor-Push Stress Tests
 - III.A.4.8.2 Maximum Loss
- III.A.4.9 Extreme-Value Theory as a Stress-Testing Method

III.A.4.10 Summary and Conclusions

B – CREDIT RISK

III.B.1 Credit Risk Management

Author to be confirmed

III.B.2 Foundations of Credit Risk Modelling

Philipp Schönbucher

III.B.2.1 Introduction

III.B.2.2 What is Default Risk?

III.B.2.3 Exposure, Default and Recovery Processes

III.B.2.4 The Credit Loss Distribution

III.B.2.5 Expected and Unexpected Loss

III.B.2.6 Recovery Rates

III.B.2.7 Conclusion

III.B.3 Credit Exposure

Philipp Schönbucher

III.B.3.1 Introduction

III.B.3.2 Pre-settlement versus Settlement Risk

III.B.3.2.1 Pre-settlement Risk

III.B.3.2.2 Settlement Risk

III.B.3.3 Exposure Profiles

III.B.3.3.1 Exposure Profiles of Standard Debt Obligations

III.B.3.3.2 Exposure Profiles of Derivatives

III.B.3.4 Mitigation of Exposures

III.B.3.4.1 Netting Agreements

III.B.3.4.2 Collateral

III.B.3.4.3 Other Counterparty Risk Mitigation Instruments

III.B.4 Default and Credit Migration

Philipp Schönbucher

III.B.4.1 Default Probabilities and Term Structures of Default Rates

III.B.4.2 Credit Ratings

III.B.4.2.1 Measuring Rating Accuracy

III.B.4.3 Agency Ratings

III.B.4.3.1 Methodology

III.B.4.3.2 Transition Matrices, Default Probabilities and Credit Migration

III.B.4.4 Credit Scoring and Internal Rating Models

III.B.4.4.1 Credit Scoring

III.B.4.4.2 Estimation of the Probability of Default

III.B.4.4.3 Other Methods to Determine the Probability of Default

III.B.4.5 Market Implied Default Probabilities

III.B.4.5.1 Pricing the Calibration Securities

III.B.4.5.2 Calculating implied default probabilities

III.B.4.6 Credit rating and credit spreads

III.B.4.7 Summary

III.B.5 Portfolio Models of Credit Loss

Michel Crouhy, Dan Galai, Robert Mark

- III.B.5.1 Introduction
- III.B.5.2 What Actually Drives Credit Risk at the Portfolio Level?
- III.B.5.3 Credit Migration Framework
 - III.B.5.3.1 Credit VaR for a Single Bond/Loan
 - III.B.5.3.2 Estimation of Default and Rating Changes Correlations
 - III.B.5.3.3 Credit VaR of a Bond/Loan Portfolio
- III.B.5.4 Conditional Transition Probabilities– CreditPortfolioView
- III.B.5.5 The Contingent Claim Approach to Measuring Credit Risk
 - III.B.5.5.1 Structural Model of Default Risk: Merton’s (1974) Model
 - III.B.5.5.2 Estimating Credit Risk as a Function of Equity Value
- III.B.5.6 The KMV Approach
 - III.B.5.6.1 Estimation of the Asset Value VA and the Volatility of Asset Return
 - III.B.5.6.2 Calculation of the ‘Distance to Default’
 - III.B.5.6.3 Derivation of the Probabilities of Default from the Distance to Default
 - III.B.5.6.4 EDF as a Predictor of Default
- III.B.5.7 The Actuarial Approach
- III.B.5.8 Summary and Conclusion

III.B.6 Credit Risk Capital Calculation

Dan Rosen

- III.B.6.1 Introduction
- III.B.6.2 Economic Credit Capital Calculation
 - III.B.6.2.1 Economic Capital and the Credit Portfolio Model
 - III.B.6.2.1.1 Time Horizon
 - III.B.6.2.1.2 Credit Loss Definition
 - III.B.6.2.1.3 Quantile of the Loss Distribution
 - III.B.6.2.2 Expected and Unexpected Losses
 - III.B.6.2.3 Enterprise Credit Capital and Risk Aggregation
- III.B.6.3 Regulatory Credit Capital: Basel I
 - III.B.6.3.1 Minimum Credit Capital Requirements under Basel I
 - III.B.6.3.2 Weaknesses of the Basel I Accord for Credit Risk
 - III.B.6.3.3 Regulatory Arbitrage
- III.B.6.4 Regulatory Credit Capital: Basel II
 - III.B.6.4.1 Latest Proposal for Minimum Credit Capital requirements
 - III.B.6.4.2 The Standardised Approach in Basel II
 - III.B.6.4.3 Internal Ratings Based Approaches: Introduction
 - III.B.6.4.4 IRB for Corporate, Bank and Sovereign Exposures
 - III.B.6.4.5 IRB for Retail Exposures
 - III.B.6.4.6 IRB for SME Exposures
 - III.B.6.4.7 IRB for Specialised Lending and Equity Exposures
 - III.B.6.4.8 Comments on Pillar II
- III.B.6.5 Basel II: Credit Model Estimation and Validation
 - III.B.6.5.1 Methodology for PD Estimation
 - III.B.6.5.2 Point-in-Time and Through-the-Cycle Ratings
 - III.B.6.5.3 Minimum Standards for Quantification and Credit Monitoring Processes
 - III.B.6.5.4 Validation of Estimates
- III.B.6.6 Basel II: Securitisation
- III.B.6.7 Advanced Topics on Economic Credit Capital
 - III.B.6.7.1 Credit Capital Allocation and Marginal Credit Risk Contributions
 - III.B.6.7.2 Shortcomings of VaR for ECC and Coherent Risk Measures
- III.B.6.8 Summary and Conclusions

C – OPERATIONAL RISK

III.C.1 The Operational Risk Management Framework

Michael Ong

- III.C.1.1 Introduction
- III.C.1.2 Evidence of Operational Failures
- III.C.1.3 Defining Operational Risk
- III.C.1.4 Types of Operational Risk
- III.C.1.5 Aims and Scope of Operational Risk Management
- III.C.1.6 Key Components of Operational Risk
- III.C.1.7 Supervisory Guidance on Operational Risk
- III.C.1.8 Identifying Operational Risk – the Risk Catalogue
- III.C.1.9 The Operational Risk Assessment Process
- III.C.1.10 The Operational Risk Control Process
- III.C.1.11 Some Final Thoughts

III.C.2 Operational Risk Process Models

James Lam

- III.C.2.1 Introduction
- III.C.2.2 The Overall Process
- III.C.2.3 Specific Tools
- III.C.2.4 Advanced Models
 - III.C.2.4.1 Top-down models
 - III.C.2.4.2 Bottom-up models
- III.C.2.5 Key Attributes of the ORM Framework
- III.C.2.6 Integrated Economic Capital Model
- III.C.2.7 Management Actions
- III.C.2.8 Risk Transfer
- III.C.2.9 IT Outsourcing
 - III.C.2.9.1 Stakeholder Objectives
 - III.C.2.9.2 Key Processes
 - III.C.2.9.3 Performance Monitoring
 - III.C.2.9.4 Risk Mitigation

III.C.3 Operational Value-at-Risk

Carol Alexander

- III.C.3.1 The ‘Loss Model’ Approach
- III.C.3.2 The Frequency Distribution
- III.C.3.3 The Severity Distribution
- III.C.3.4 The Internal Measurement Approach
- III.C.3.5 The Loss Distribution Approach
- III.C.3.6 Aggregating ORC
- III.C.3.7 Concluding Remarks

Introduction

If you're reading this, you are seeking to attain a higher standard. Congratulations!

Those of us who have been a part of financial risk management for the past twenty years, have seen it change from an on-the-fly profession, with improvisation as a rule, to one with substantially higher standards, many of which are now documented and expected to be followed. It's no longer enough to *say* you know. Now, you and your team need to *prove* it.

As its title implies, this book is *the* Handbook for the Professional Risk Manager. It is for those professionals who seek to demonstrate their skills through certification as a Professional Risk Manager (PRM) in the field of financial risk management. And it is for those looking simply to develop their skills through an excellent reference source.

With contributions from nearly 40 leading authors, the Handbook is designed to provide you with the materials needed to gain the knowledge and understanding of the building blocks of professional financial risk management. Financial risk management is not about avoiding risk. Rather, it is about understanding and communicating risk, so that risk can be taken more confidently and in a better way. Whether your specialism is in insurance, banking, energy, asset management, weather, or one of myriad other industries, this Handbook is your guide.

We encourage you to work through it sequentially. In Section I, we introduce the foundations of finance theory, the financial instruments that provide tools for the mitigation or transfer of risk, and the financial markets in which instruments are traded and capital is raised. After studying this section, you will have read the materials necessary for passing Exam I of the PRM Certification program.

In Section II, we take you through the mathematical foundations of risk assessment. While there are many nuances to the practice of risk management that go beyond the quantitative, it is essential today for every risk manager to be able to assess risks. The chapters in this section are accessible to all PRM members, including those without any quantitative skills. The Excel spreadsheets that accompany the examples are an invaluable aid to understanding the mathematical and statistical concepts that form the basis of risk assessment. After studying all these chapters, you will have read the materials necessary for passage of Exam II of the PRM Certification program.

In Section III, the current and best practices of Market, Credit and Operational risk management are described. This is where we take the foundations of Sections I and II and apply them to our

profession in very specific ways. Here the strategic application of risk management to capital allocation and risk-adjusted performance measurement takes hold. After studying this part, you will have read the materials necessary for passage of Exam III of the PRM Certification program.

Those preparing for the PRM certification will also be preparing for Exam IV - Case Studies, Standards of Best Practice Conduct and Ethics and PRMIA Governance. This is where we study some failed practices, standards for the performance of the duties of a Professional Risk Manager, and the governance structure of our association, the Professional Risk Managers' International Association. The materials for this exam are freely available on our web site (see http://www.prmia.org/pdf/Web_based_Resources.htm) and are thus outside of the Handbook.

At the end of your progression through these materials, you will find that you have broadened your knowledge and skills in ways that you might not have imagined. You will have challenged yourself as well. And, you will be a better risk manager. It is for this reason that we have created the Professional Risk Managers' Handbook.

Our deepest appreciation is extended to Prof. Carol Alexander and Prof. Elizabeth Sheedy, both of PRMIA's Academic Advisory Council, for their editorial work on this document. The commitment they have shown to ensuring the highest level of quality and relevance is beyond description. Our thanks also go to Laura Bianco, President of PRMIA Publications, who has tirelessly kept the work process moving forward and who has dedicated herself to demanding the finest quality output. We also thank Richard Leigh, our London-based copyeditor, for his skilful and timely work.

Finally, we express our thanks to the authors who have shared their insights with us. The demands for sharing of their expertise are frequent. Yet, they have each taken special time for this project and have dedicated themselves to making the Handbook and *you* a success. We are very proud to bring you such a fine assembly.

Much like PRMIA, the Handbook is a place where the best ideas of the risk profession meet. We hope that you will take these ideas, put them into practice and certify your knowledge by attaining the PRM designation. Among our membership are over 300 Chief Risk Officers / Heads of Risk and 800 other senior executives who will note your achievements. They too know the importance of setting *high standards* and the trust that capital providers and stakeholders have put in them. Now they put their trust in you and you can prove your commitment and distinction to them.

We wish you much success during your studies and for your performance in the PRM exams!

David R. Koenig, Executive Director, PRMIA

Section I

Finance Theory, Financial Instruments and Markets

Preface to Section I: Finance Theory, Financial Instruments and Markets

Section I of this Handbook has been written by a group of leading scholars and practitioners and represents a broad overview of the theory, instruments and markets of finance. This section corresponds to Exam I in the Professional Risk Manager (PRM) certification programme.

The modern theory of finance is the solid basis of risk management and thus it naturally represents the basis of the PRM certification programme. All major areas of finance are involved in the process of risk management: from the expected utility approach and risk aversion, which were the forerunners of the capital asset pricing model (CAPM), to portfolio theory and the risk-neutral approach to pricing derivatives. All of these great financial theories and their interactions are presented in Part I.A (Finance Theory). Many examples demonstrate how the concepts are applied in practical situations.

Part I.B (Financial Instruments) describes a wide variety of financial products and connects them to the theoretical development in Part I.A. The ability to value all the instruments/assets within a trading or asset portfolio is fundamental to risk management. This part examines the valuation of financial instruments and also explains how many of them can be used for risk management.

The designers of the PRM curriculum have correctly determined that financial risk managers should have a sound knowledge of financial markets. Market liquidity, the role of intermediaries and the role of exchanges are all features that vary considerably from one market to the next and over time. It is crucial that professional risk managers understand how these features vary and their consequences for the practice of risk management. Part I.C (Financial Markets) describes where and how instruments are traded, the features of each type of financial asset or commodity and the various conventions and rules governing their trade.

This background is absolutely necessary for professional risk management, and Exam I therefore represents a significant portion of the whole PRM certification programme. For a practitioner who left academic studies several years ago, this part of the Handbook will provide efficient revision of finance theory, financial instruments and markets, with emphasis on practical application to risk management. Such a person will find the chapters related to his/her work easy reading and will have to study other topics more deeply.

The coverage of financial topics included in Section I of the Handbook is typically deeper and broader than that of a standard MBA syllabus. But the concepts are well explained and

appropriately linked together. For example, Chapter I.B.6 on credit derivatives covers many examples (such as credit-linked notes and credit default swaps) that are not always included in a standard MBA-level elective course on fixed income. Chapter I.B.9 on simple exotics also provides examples of path-dependent derivatives beyond the scope of a standard course on options. All chapters are written for professionals and assume a basic understanding of markets and their participants.

Finance Theory

Chapter I.A.1 provides a general overview of risk and risk aversion, introduces the utility function and mean–variance criteria. Various risk-adjusted performance measures are described. A summary of several widely used utility functions is presented in the appendix.

Chapter I.A.2 provides an introduction to portfolio mathematics, from means and variances of returns to correlation and portfolio variance. This leads the reader to the efficient frontier, portfolio theory and the concept of portfolio diversification. Eventually this chapter discusses normally distributed returns and basic applications for value-at-risk, as well as the probability of reaching a target or beating a benchmark. This chapter is very useful for anybody with little experience in applying basic mathematical models in finance.

The concept of capital allocation is another fundamental notion for risk managers. Chapter I.A.3 describes how capital is allocated between portfolios of risky and riskless assets, depending on risk preference. Then the efficient frontier, the capital markets line, the Sharpe ratio and the separation principle are introduced. These concepts lead naturally to a discussion of the CAPM model and the idea that marginal risk (rather than absolute risk) is the key issue when pricing risky assets. Chapter I.A.4 provides a rigorous description of the CAPM model, including betas, systematic risk, alphas and performance measures. Arbitrage pricing theory and multifactor models are also introduced in this chapter.

Capital structure is an important theoretical concept for risk managers since capital is viewed as the last defence against extreme, unexpected outcomes. Chapter I.A.5 introduces capital structure, advantages and costs related to debt financing, various agency costs, various types of debt and equity, return on equity decomposition, examples of attractive and unattractive debt, bankruptcy and financial distress costs.

Most valuation problems involve discounting future cash flows, a process that requires knowledge of the term structure of interest rates. Chapter I.A.6 describes various types of

interest rates and discounting, defines the term structure of interest rates, introduces forward rates and explains the three main economic term structure theories.

These days all risk managers must be well versed in the use and valuation of derivatives. The two basic types of derivatives are forwards (having a linear payoff) and options (having a non-linear payoff). All other derivatives can be decomposed to these underlying payoffs or alternatively they are variations on these basic ideas. Chapter I.A.7 describes valuation methods used for forward contracts. Discounting is used to value forward contracts with and without intermediate cash flow. Chapter I.A.8 introduces the principles of option pricing. It starts with definitions of basic put and call options, put–call parity, binomial models, risk-neutral methods and simple delta hedging. Then the Black–Scholes–Merton formula is introduced. Finally, implied volatility and smile effects are briefly described.

Financial Instruments

Having firmly established the theoretical basis for valuation in Part I.A, Part I.B applies these theories to the most commonly used financial instruments.

Chapter I.B.1 introduces bonds, defines the main types of bonds and describes the market conventions for major types of treasuries, strips, floaters (floating-rate notes) and inflation-protected bonds in different countries. Bloomberg screens are used to show how the market information is presented. Chapter I.B.2 analyses the main types of bonds, describes typical cash flows and other features of bonds and also gives a brief description of non-conventional instruments. Examples of discounting, day conventions and accrued interest are provided, as well as yield calculations. The connection between yield and price is described, thus naturally leading the reader to duration, convexity and hedging interest-rate risk.

While Chapter I.A.7 explained the principles of forward valuation, Chapter I.B.3 examines and compares futures and forward contracts. Usage of these contracts for hedging and speculation is discussed. Examples of currency, commodity, bonds and interest-rate contracts are used to explain the concept and its applications. Mark-to-market, quotation, settlements and other specifications are described here as well. The principles of forward valuation are next applied to swap contracts, which may be considered to be bundles of forward contracts. Chapter I.B.4 analyses some of the most popular swap varieties, explaining how they may be priced and used for managing risk.

The remaining chapters in Part I.B all apply the principles of option valuation as introduced in Chapter I.A.8. The power of the option concept is obvious when we see its applications to so

many instruments and risk management problems. Chapter I.B.5 begins with an analysis of vanilla options. Chapter I.B.6 covers one of the newer applications of options: the use of credit risk derivatives to manage credit risk. Chapter I.B.7 addresses caps, floors and swaptions, which are the main option strategies used in interest-rate markets. Yet another application of the option principle is found in Chapter I.B.8 – convertible bonds. These give investors the right to convert a debt security into equity. Finally, Chapter I.B.9 examines exotic option payoffs. In every case the author defines the instrument, discusses its pricing and illustrates its use for risk management purposes.

Financial Markets

Financial risk management takes place in the context of markets and varies depending on the nature of the market. Chapter I.C.1 is a general introduction to world financial markets. They can be variously classified – geographically, by type of exchange, by issuers, liquidity and type of instruments – all are provided here. The importance of liquidity, the distinction between exchange and over-the-counter markets and the role of intermediaries in their various forms are explained in more detail.

Money markets are the subject of Chapter I.C.2. These markets are of vital importance to the risk manager as the closest thing to a ‘risk-free’ asset is found here. This chapter covers all short-term debt securities, whether issued by governments or corporations. It also explains the repo markets – markets for borrowing/lending on a secured basis. The market for longer-term debt securities is discussed in Chapter I.C.3, which classifies bonds by issuer: government, agencies, corporate and municipal. There is a comparison of bond markets in major countries and a description of the main intermediaries and their roles. International bond markets are introduced as well.

Chapter I.C.4 turns to the foreign exchange market – the market with the biggest volume of trade. Various aspects of this market are explained, such as quotation conventions, types of brokers, and examples of cross rates. Economic theories of exchange rates are briefly presented here along with central banks’ policies. Forward rates are introduced together with currency swaps. Interest-rate parity is explained with several useful examples.

Chapter I.C.5 provides a broad introduction to stock markets. This includes the description and characteristics of several types of stocks, stock market indices and priorities in the case of liquidation. Dividends and dividend-based stock valuation methods are described in this chapter. Primary and secondary markets are distinguished. Market mechanics, including types of

orders, market participants, margin and short trades, are explained here with various examples clarifying these transactions. Some exchange-traded options on stocks are introduced as well.

Chapter I.C.6 introduces the futures markets; this includes a comparison of the main exchange-traded markets, options on futures, specifications of the most popular contracts, the use of futures for hedging, trade orders for futures contracts, mark-to-market procedures, and various expiration conventions. A very interesting description of the main market participants concludes this chapter.

Chapter I.C.7 introduces the structure of the commodities market. It starts with the spot market and then moves to commodity forwards and futures. Specific features, such as delivery and settlement methods, are described. The spot–forward pricing relationship is used to decompose the forward price into spot and carry. Various types of price term structure (such as backwardation and contango) are described, together with some economic theory. The chapter also describes short squeezes and regulations. Risk management at the commodity trading desk is given at a good intuitive level. The chapter concludes with some interesting facts on distribution of commodity returns.

Finally, Chapter I.C.8 examines one of the most rapidly developing markets for risk – the energy markets. These markets allow participants to manage the price risks of oil and gas, electricity, coal and so forth. Some other markets closely linked with energy are also briefly discussed here, including markets for greenhouse gas emissions, weather derivatives and freight. Energy markets create enormous challenges and opportunities for risk managers – in part because of the extreme volatility of prices that can occur.

As a whole, Section I gives an overview of the theoretical and practical aspects of finance that are used in the management of financial risks. Many concepts, some quite complex, are explained in a relatively simple language and are demonstrated with numerous examples. Studying this part of the Handbook should refresh your knowledge of financial models, products and markets and provide the background for risk management applications.

Zvi Wiener, Co-chair of PRMLA's Education and Standards Committee

Section II

Mathematical Foundations of Risk Measurement

Preface to Section II: Mathematical Foundations of Risk Measurement

The role of risk management in financial firms has evolved far beyond the simple insurance of identified risks. Today it is recognised that risks cannot be properly managed unless they are quantified. And the assessment of risk requires mathematics. Take, for instance, a large portfolio of stocks. The relationship between the portfolio returns and the market returns – and indeed other potential risk factor returns – is typically estimated using a statistical regression analysis. And the systematic risk of the portfolio is then determined by a quadratic form, a fundamental concept in matrix algebra that is based on the covariance matrix of the risk factor returns.

Volatility is not the only risk metric that financial risk managers need to understand. During the last decade value-at-risk (VaR) has become the ubiquitous tool for risk capital estimation. To understand a VaR model, risk managers require knowledge of probability distributions, simulation methods and a host of other mathematical and statistical techniques. Market VaR is assessed by mapping portfolios to their risk factors and forecasting the volatilities and correlations of these factors. The diverse quantitative techniques that are commonly applied in the assessment of market VaR include eigenvectors and eigenvalues, Taylor expansions and partial derivatives. Credit VaR can be assessed using firm-value models that are based on the theory of options, or statistical and/or macro-econometric models. Probability distributions are even applied to operational risks, though they are very difficult to quantify because the data are sparse and unreliable. Indeed, the actuarial or loss model approach has been adopted as industry standard for operational VaR models.

Even if not directly responsible for designing and coding a risk capital model, middle office risk managers must understand these models sufficiently well to be competent to assess them. And the risk management role encompasses many other responsibilities. Ten years ago my best students aspired to become traders because of the high salaries and status – risk management was viewed (by some) as a ‘second-rate’ job that did not require very special expertise. Now, this situation has definitely changed. Today, the middle office risk manager’s responsibility has expanded to include the independent validation of traders’ models, as well as risk capital assessment. And the role of risk management in the front office itself has expanded, with the need to hedge increasingly complex options portfolios. So today, the hallmark of a good risk manager is not just having the statistical skills required for risk assessment – a comprehensive knowledge of pricing and hedging financial instruments is equally important.

No wonder, therefore, that the PRM qualification includes an entire exam on mathematical and statistical methods. However, we do recognise that many students will not have degrees in mathematics, physics or other quantitative disciplines. So this section of the Handbook is aimed at students having no quantitative background at all. It introduces and explains all the mathematics and statistics that are essential for financial risk management. Every chapter is presented in a pedagogical manner, with associated Excel spreadsheets explaining the numerous practical examples. And, for clarity and consistency, we chose two much respected authors of the highly acclaimed textbook *Quantitative Methods in Finance* to write the entire section. Keith Parramore and Terry Watsham have put considerable effort into making the PRM material accessible to everyone, irrespective of their quantitative background.

The first chapter, II.A (Foundations), reviews the fundamental mathematical concepts: the symbols used and the basic rules for arithmetic, equations and inequalities, functions and graphs, etc. Chapter II.B (Descriptive Statistics) introduces the descriptive statistics that are commonly used to summarise the historical characteristics of financial data: the sample moments of returns distributions, ‘downside’ risk statistics, and measures of covariation (e.g. correlation) between two random variables. Chapter II.C (Calculus) focuses on differentiation and integration, Taylor expansion and optimisation. Financial applications include calculating the convexity of a bond portfolio and the estimation of the delta and gamma of an options portfolio. Chapter II.D (Linear Mathematics and Matrix Algebra) covers matrix operations, special types of matrices and the laws of matrix algebra, the Cholesky decomposition of a matrix, and eigenvalues and eigenvectors. Examples of financial applications include: manipulating covariance matrices; calculating the variance of the returns to a portfolio of assets; hedging a vanilla option position; and simulating correlated sets of returns. Chapter II.E (Probability Theory) first introduces the concept of probability and the rules that govern it. Then some common probability distributions for discrete and continuous random variables are described, along with their expectation and variance and various concepts relating to joint distributions, such as covariance and correlation, and the expected value and variance of a linear combination of random variables. Chapter II.F (Regression Analysis) covers the simple and multiple regression models, with applications to the capital asset pricing model and arbitrage pricing theory. The statistical inference section deals with both prediction and hypothesis testing, for instance, of the efficient market hypothesis. Finally, Chapter II.G (Numerical Methods) looks at solving implicit equations (e.g. the Black–Scholes formula for implied volatility), lattice methods, finite differences and simulation. Financial applications include option valuation and estimating the ‘Greeks’ for complex options.

Whilst the risk management profession is no doubt becoming increasingly quantitative, the quantification of risk will never be a substitute for good risk management. The primary role of a

financial risk manager will always be to understand the markets, the mechanisms and the instruments traded. Mathematics and statistics are only tools, but they are necessary tools. After working through this part of the Handbook you will have gained a thorough and complete grounding in the essential quantitative methods for your profession.

Carol Alexander, Chair of PRMLA's Academic Advisory Council and co-editor of The PRM Handbook

Section III

Risk Management Practices

Preface to Section III: Risk Management Practices

Section III is the ultimate part of *The PRM Handbook* in both senses of the word. Not only is it the final section, but it represents the final aims and objectives of the Handbook. Sections I (Finance Theory, Financial Instruments and Markets) and II (Mathematical Foundations of Risk Measurement) laid the necessary foundations for this discussion of risk management practices – the primary concern of most readers. Here some of the foremost practitioners and academics in the field provide an up-to-date, rigorous and lucid statement of modern risk management.

The practice of risk management is evolving at a rapid pace, especially with the impending arrival of Basel II. Aside from these regulatory pressures, shareholders and other stakeholders increasingly demand higher standards of risk management and disclosure of risk. In fact, it would not be an overstatement to say that risk consciousness is one of the defining features of modern business. Nowhere is this truer than in the financial services industry. Interest in risk management is at an unprecedented level as institutions gather data, upgrade their models and systems, train their staff, review their remuneration systems, adapt their business practices and scrutinise controls for this new era.

Section III is itself split into three parts which address market risk, credit risk and operational risk in turn. These three are the main components of risk borne by any organisation, although the relative importance of the mix varies. For a traditional commercial bank, credit risk has always been the most significant. It is defined as the risk of default on debt, swap, or other counterparty instruments. Credit risk may also result from a change in the value of a security, contract or asset resulting from a change in the counterparty's creditworthiness. In contrast, market risk refers to changes in the values of securities, contracts or assets resulting from movements in exchange rates, interest rates, commodity prices, stock prices, etc. Operational risk, the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events, is not, strictly speaking, a financial risk. Operational risks are, however, an inevitable consequence of any business undertaking. For financial institutions and fund managers, credit and market risks are taken intentionally with the objective of earning returns, while operational risks are a by-product to be controlled. While the importance of operational risk management is increasingly accepted, it will probably never have the same status in the finance industry as credit and market risk which are the chosen areas of competence.

For non-financial firms, the priorities are reversed. The focus should be on the risks associated with the particular business; the production and marketing of the service or product in which

expertise is held. Market and credit risks are usually of secondary importance as they are a by-product of the main business agenda.

The last line of defence against risk is capital, as it ensures that a firm can continue as a going concern even if substantial and unexpected losses are incurred. Accordingly, one of the major themes of Section III is how to determine the appropriate size of this capital buffer. How much capital is enough to withstand unusual losses in each of the three areas of risk? The measurement of risk has further important implications for risk management as it is increasingly incorporated into the performance evaluation process. Since resources are allocated and bonuses paid on the basis of performance measures, it is essential that they be appropriately adjusted for risk. Only then will appropriate incentives be created for behaviour that is beneficial for shareholders and other stakeholders. Chapter III.0 explores this fundamental idea at a general level, since it is relevant for each of the three risk areas that follow.

Market Risk

Chapter III.A.1 introduces the topic of market risk as it is practised by bankers, fund managers and corporate treasurers. It explains the four major tasks of risk management (identification, assessment, monitoring and control/mitigation), thus setting the scene for the quantitative chapters that follow. These days one of the major tasks of risk managers is to measure risk using value-at-risk (VaR) models. VaR models for market risk come in many varieties. The more basic VaR models are the topic of Chapter III.A.2, while the advanced versions are covered in III.A.3 along with some other advanced topics such as risk decomposition. The main challenge for risk managers is to model the empirical characteristics observed in the market, especially volatility clustering. The advanced models are generally more successful in this regard, although the basic versions are easier to implement. Realistically, there will never be a perfect VaR model, which is one of the reasons why stress tests are a popular tool. They can be considered an ad hoc solution to the problem of model risk. Chapter III.A.4 explains the need for stress tests and how they might usefully be constructed.

Credit Risk

Chapter III.B.1 introduces the sphere of credit risk management. Some fundamental tools for managing credit risk are explained here, including the use of collateral, credit limits and credit derivatives. Subsequent chapters on credit risk focus primarily on its modelling. Foundations for modelling are laid in Chapter III.B.2, which explains the three basic components of a credit loss: the exposure, the default probability and the recovery rate. The product of these three, which can be defined as random processes, is the credit loss distribution. Chapter III.B.3 takes a more detailed look at the exposure amount. While relatively simple to define for standard loans,

assessment of the exposure amount can present challenges for other credit sensitive instruments such as derivatives, whose values are a function of market movements. Chapter III.B.4 examines in detail the default probability and how it can evolve over time. It also discusses the relationship between credit ratings and credit spreads, and credit scoring models. Chapter III.B.5 tackles one of the most crucial issues for credit risk modelling: how to model credit risk in a portfolio context and thereby estimate credit VaR. Since diversification is one of the most important tools for the management of credit risk, risk measures on a portfolio basis are fundamental. A number of tools are examined, including the credit migration approach, the contingent claim or structural approach, and the actuarial approach. Finally, Chapter III.B.6 extends the discussion of credit VaR models to examine credit risk capital. It compares both economic capital and regulatory capital for credit risk as defined under the new Basel Accord.

Operational Risk

The framework for managing operational risk is first established in Chapter III.C.1. After defining operational risk, it explains how it may be identified, assessed and controlled. Chapter III.C.2 builds on this with a discussion of operational risk process models. By better understanding business processes we can find the sources of risk and often take steps to re-engineer these processes for greater efficiency and lower risk. One of the most perplexing issues for risk managers is to determine appropriate capital buffers for operational risks. Operational VaR is the subject of Chapter III.C.3, including discussion of loss models, standard functional forms, both analytical and simulation methods, and the aggregation of operational risk over all business lines and event types.

Elizabeth Sheedy, Member of PRMLA's Academic Advisory Council and co-editor of The PRM Handbook

I.A.1 Risk and Risk Aversion

Jacques Pézier¹

I.A.1.1 Introduction

Risk management, in a wide sense, is the art of making decisions in an uncertain world. Such decisions involve a weighting of risks and rewards, a choice between doing the safe thing and taking a risk. For example, we may ponder whether to invest in a new venture, whether to diversify or hedge a portfolio of assets, or at what price it would be worth insuring a person or a system. Risk attitude determines such decisions. Utility theory offers a rational method for expressing risk attitude and should therefore be regarded as a main pillar of risk management. The other two pillars of risk management are the generation of good alternatives – without which there would be nothing to decide – and the assessment of probabilities – without which we could not tell the likely consequences of our actions.

Rationality, in the context of utility theory, means simply that decisions should be logically consistent with a set of preference axioms and in line with patterns of risk attitude expressed in simple, easily understood circumstances. So, utility theory does not dictate what risk attitude should be – that remains a personal matter or a matter of company policy – it merely provides a logical framework to extend risk preferences from simple cases to complex situations.

But why should one seek an axiomatic framework to express risk preferences? Alas, experience shows that unaided intuition is an unreliable guide. It is relatively easy to construct simple decision problems where intuitive choices seem to contradict each other, that is, seem to violate basic rules of behaviour that we hold as self-evident. It seems wise, therefore, to start by agreeing a basic set of rules and then draw the logical consequences.

Thus, utility theory is neither purely descriptive nor purely normative. It brings about a more disciplined, quantitative approach to the expression of risk attitude than is commonly found in everyday life. In other words, where too often risk taking is ‘seat of the pants’ or based on ‘gut feel’ or ‘nose’, it tries to bring the brain into play. By questioning instinctive reactions to risky situations, it leads decision makers and firms to understand better what risk attitude they ought to adopt, to express it formally as an element of corporate policy and to convey it through the organisation so that decisions under uncertainty can be safely delegated.

¹ Visiting Professor, ISMA Centre, University of Reading, UK.

This chapter introduces some concepts that are absolutely fundamental to the management of financial risks. Section I.A.1.2 introduces the idea of utility maximisation following Bernoulli's original ideas. Section I.A.1.3 discusses the 'axiom of independence of choice', one of the basic axioms that must be satisfied if preferences over risky outcomes are to be represented by a utility function. Section I.A.1.4 introduces the principle of maximum expected utility and the concept of risk aversion (and its inverse, risk tolerance). Section I.A.1.5 explains how to encode your personal attitude to risk in your own utility function. Section I.A.1.6 shows under what circumstances the principle of maximum expected utility reduces to a mean–variance criterion to distinguish between different investments. A comprehensive treatment of risk-adjusted performance measures is given in Section I.A.1.7. We pay particular attention to the circumstances in which the risks to be compared are not normally distributed and investors are mainly concerned with downside risks. Section I.A.1.8 summarises and indicates which types of decision criteria and performance measures may be appropriate in which circumstances.

Much of the material that is introduced in this chapter will be more fully discussed in other parts of the *Handbook*. Thus you will find many references to subsequent chapters in Part I.A, Part II and Part III of the *Handbook*. A thorough treatment of utility theory, whilst fundamental to our understanding of *risk and risk aversion*, is beyond the scope of the PRM exam. However, for completeness, and for readers seeking to use this chapter as a resource that goes further than the PRM syllabus, we have provided extensive footnotes of the mathematical derivations. Furthermore, we have added an Appendix that describes the properties of standard utility functions. However, it should be stressed that neither the mathematical derivations in the footnotes nor the material in Appendix B are part of the PRM exam.

I.A.1.2 Mathematical Expectations: Prices or Utilities?

It may seem curious nowadays that early probabilists, who liked to study games of chance, took it for granted that the mathematical expectation of cash outcomes was the only rational criterion for choosing between gambles. The *expected value* of a gamble is defined as the sum of its cash outcomes weighted by their respective probabilities; the gamble with the highest expected value was deemed to be the best. Fairness in gambling was the main argument in support of this principle (among 'zero-sum' games, where the gains of one player are the losses of the other, only zero-expectation games are fair). Another argument drew on the *weak law of large numbers*, which implies that, if the consequences of each gamble are small relative to the wealth of the players, then, in the long run, after many independent gambles, only the average result would matter.

Daniel Bernoulli (1738) was the first mathematician to question the principle of maximising expected value and to try to justify departures from it observed in daily life. He questioned

choices that fly in the face of the principle of maximising expected value. For example, he asked, if a poor man were offered an equal chance to win a fortune or nothing, should he be regarded as irrational if he tried to negotiate a sure reward of slightly less than half the potential fortune? Or is it insane to insure a precious asset and thus knowingly contribute an expected profit to the insurance company and therefore an equivalent expected decrease in one's wealth? To reconcile common behaviour with a maximum-expectation principle, Bernoulli suggested applying the principle not to cash outcomes but to utilities² associated with cash outcomes. Bernoulli thus pre-dates by half a century the core tenet of the Utilitarianism school of social philosophy, the distinction between:

- the *utility*, i.e. the 'personal value' of an asset,
- and
- the *price*, i.e. the 'exchange value' of an asset.

Bernoulli's principle was that actions should be directed at *maximising expected utility*. The problem that inspired Bernoulli and which has gained fame under the name of the *St Petersburg paradox*³ runs as follows: Peter tosses a coin and continues to do so until it should land 'heads'. He agrees to give Paul one ducat if he gets 'heads' on the very first throw, two ducats if he gets it on the second, four if on the third, eight if on the fourth, and so on, so that with each additional throw the number of ducats he must pay is doubled. We seek to determine the value of Paul's expectation.

Since, with a fairly tossed, symmetrical coin, the probability of landing heads for the first time on the k th toss is 2^{-k} and the corresponding reward is $2^k - 1$ ducats, the contribution to Paul's monetary expectation of this outcome is half a ducat. And since there is an infinite number of possible outcomes $k = 1, k = 2, \text{ etc.}$, Paul's monetary expectation is infinite. But, then as now, gamblers are not willing to pay more than a few ducats for the right to play the game, hence the paradox.

Bernoulli suggested that the utility of a cash reward depends on the existing wealth of the recipient. He even made the far stronger assumption that *utility is always inversely proportional to existing wealth*, in other words, that a gain of *one* ducat to someone worth a thousand ducats has the same utility as a gain of *a thousand* ducats to someone worth a million ducats.

In this case a small change in utility, du , would be related to a small change in wealth, dx , by

$$du = dx/x.$$

² In the Latin original, to calculate an 'emolumentum medium'.

³ Simply because Bernoulli's paper was published in the *Commentaries from the Academy of Sciences of St Petersburg*.

This leads, by integration (see Section II.C.6), to a *logarithmic* utility function,

$$u(x) = \ln(x).$$

If we apply a logarithmic utility to the St Petersburg paradox, it no longer appears to be a paradox. For instance, a gambler whose only wealth is the game itself would perceive an expected utility of $\ln(2)$, which is the same as the utility of 2 ducats, and this is quite a small number – far short of infinity! In general, if the gambler has a logarithmic utility function, the larger the initial wealth of the gambler, the larger his perceived utility of the game.

I.A.1.3 The Axiom of Independence of Choice

Rarely is the power of a new idea fully understood on first encounter. Bernoulli's introduction of a utility function *did* influence the development of classical economics, where it was transposed into a deterministic context. But it took more than two hundred years for the concept to be revived in its original probabilistic context and to be re-erected on a firmer footing. In a seminal book on games theory the mathematician J. von Neumann and the economist O. Morgenstern (1947) postulated a basic set of rules from which it will follow that a utility function provides a *complete* description of an individual's risk attitude.⁴

Bernoulli made a very strong assumption – that the utility of a gain is inversely proportional to existing wealth. By contrast, von Neumann and Morgenstern only assumed a minimal set of rules that should appeal to all decision makers and which would result in the existence of utilities without specifying what these utilities will be. These rules, or 'preference axioms', should seem so fundamental that if, in some circumstances, a decision maker accidentally violates one of them, she would re-examine her choice and correct it rather than knowingly abuse one of the rules.

To illustrate this point, consider the following preference axiom:

*'A choice between two gambles should not be influenced by the way the gambles are presented, provided that all presentations contain the same relevant information.'*⁵

This is an axiom because it cannot be derived from more fundamental principles. It is called the *axiom of independence of choice*. One is free to accept or reject it, though most decision makers freely accept it as self-evident. However, this axiom is easily violated by instinctive choices.

⁴ The axiomatic approach pioneered by von Neumann, Morgenstern, Savage and others is often referred to as the American school of axiomatic utility theory.

Daniel Kahneman, a Nobel prize winning expert in cognitive psychology, and his long time colleague Amos Tversky designed the following simple, if somewhat dramatic test to show how a change of presentation *can* affect our decisions. Their test consists of presenting two variants of a choice between two public health programmes that address a threat to the lives of 600 people. The first variant is:

‘With programme A we know that 200 lives will be saved, whereas with programme B there is a one-third chance of saving all 600 lives and a two-thirds chance of saving none.’

Kahneman and Tversky found that a clear majority of the people they presented with this choice preferred A to B.⁵ The second variant is:

‘With programme C we know that 400 lives will be lost, whereas with programme D there is a one-third chance that none will die and a two-thirds chance that all 600 people will die.’

A majority of the people presented with this choice prefer D to C. Now, looking at the four programmes, it becomes clear that, on the one hand, A and C are the same and, on the other hand, C and D are also the same; the people saved in one presentation are the people not dying in the other. So, whether one prefers A to B or the reverse, one ought express the same order of preference between C and D, and that is not the case with many of the people interviewed; these people are violating the axiom of independence of choice.

Kahneman and Tversky (1979) developed a new theory to explain their findings. They suggested that people are generally risk averse when choosing between a sure gain and a chance of a larger gain, but the same people may take a chance when forced to choose between a sure loss and only a probability of a worse loss. The snag is that what appears as a sure gain or a sure loss is often a question of *perspective* that can be easily manipulated by the way a problem is presented. Aware of the importance attached to presentation, we provide in Appendix I.A.1.A a brief glossary of some of the terms used in this chapter in order to dispel any unintended meaning.

More generally, cognitive psychologists have shown that we, as decision makers, may be swayed by cognitive biases in the same way as untrained observers may be tricked by optical illusions. We recognise the possibility of such biases when dealing with unusual events, for example, rare events or extreme circumstances, or when our thoughts are too accustomed to a status quo, or when they are blurred by emotions. But it may be unsafe to dismiss all instinctive reactions as mere ‘biases’. After all, human instincts have evolved over millennia and must have some

⁵ Similar hypothetical questions were presented to numerous audiences of students and university faculty (the Hebrew University of Jerusalem, University of Stockholm, University of Michigan, among others) with similar results and repeated with business men in National Science Foundation sponsored studies.

survival value; important features of human risk behaviour could be overlooked by a naïve axiomatic approach.

I.A.1.4 Maximising Expected Utility

There are a few variations of the axiomatic formulation of utility theory. We give here an intuitive, if less than rigorous presentation.⁶ For students' information we give in a footnote the derivation of the principle of maximum expected utility, but the derivation is *not* examinable in the PRM.

I.A.1.4.1 The Four Basic Axioms

(A1) *Transitivity of Choice*: All possible outcomes of the decision under consideration can be ranked in order of preference; that is, if among three outcomes A , B and C , we strictly prefer A to B and B to C then we ought to strictly prefer A to C .

(A2) *Continuity of Choice*: If among three outcomes A , B , C we strictly prefer A to B and B to C , then B is the *certain equivalent* of some lottery between A and C , that is, there exists a unique probability p for which we should be indifferent between receiving B or playing a lottery offering A with probability p and C with probability $1 - p$.

(A3) *Independence of Choice*:⁷ Our preference order between two lotteries should not be affected if these lotteries are part of the same wider range of possibilities.

(A4) *Stochastic Dominance*: Between two lotteries offering the same two possible outcomes, we ought to prefer the lottery offering the larger probability of yielding the preferred outcome.

Whether these axioms are naïve or reasonable will remain an open debate; they are certainly not always descriptive of intuitive human behaviour – see Allais (1953) as well as Khaneman – but they may be useful guides as we try to improve on intuition. What is remarkable is that these four axioms are sufficient to establish the concept of utility and lead to a unique decision criterion known as the *principle of maximum expected utility* (maximum EU, for short), namely: the lottery with the largest expected utility ought to be preferred over others.⁸

⁶ For the original presentation, see von Neumann and Morgenstern (1947). For alternative presentations, see Savage (1954), Fishburn (1970) or Kreps (1988).

⁷ The axiom of independence of choice has been formulated in many ways. In this form, it is also known as the axiom of *substitution* or simply of *no fun in gambling*.

⁸ Suppose we face a choice between two lotteries A and B , each offering some of a finite number of outcomes $\{x_i\}$, $i = 1$ to n . We associate probability p_{Ai} to outcome x_i in lottery A and p_{Bi} in lottery B , respectively. We seek a criterion that will transform the choice between the two lotteries into determining which of two real numbers is the largest. Axiom (A1) requires that we should be able to rank all outcomes in a simple preference order and therefore that we should be able to identify at least one outcome that is not less desirable than any other, call it M , and at least one outcome that is not more desirable than any other, call it m . Axiom (A2) implies that to any outcome x_i corresponds a probability u_i such that x_i can be regarded as the certain equivalent of a lottery offering M with probability u_i and m with probability $1 - u_i$. Now, for each prize offered in lotteries A and B , substitute the equivalent lottery between M and m . According to the axiom of independence of choice (A3), our preference between the new, compounded lotteries, call them A' and

Any decision criterion other than ‘maximum EU’ that leads to a different choice would violate at least one of the four basic axioms. It is therefore somewhat mystifying that the maximum EU principle is not routinely used in risk management. We address this paradox in Section I.A.1.9.

I.A.1.4.2 Introducing the Utility Function

Assigning utilities to possible outcomes is the key. We explain how this may be done in the next section. But let us remark first that, for most financial risks, outcomes are already expressed on a *monetary* scale, for instance, company profit or shareholder value. That is no mean feat and one can only hope that there is no significant loss of information or distortion in the translation process. Outcomes are generally complex, multi-faceted, and perceived differently by various interested parties: shareholders, investors, clients, employees, management, etc. We must be confident that between two outcomes A and B we prefer A to B simply because the cash value of A is greater than the cash value of B .

Utility theory does not require the expression of all outcomes on a monetary scale and therefore can address more general decision problems. However, when outcomes are already expressed in terms of cash, utilities become a function of cash; we limit our discussion to this case.

The utility function $u(x)$, where x is a cash amount expressing wealth and $u(x)$ its utility to the owner of the wealth, should be a continuous, non-decreasing function of x . It should be continuous in as much as cash itself can be considered as continuous and a small increase in cash should produce small increase in utility.⁹ It should be non-decreasing in as much as more cash is preferred to less, a proposition that is not necessarily obvious and that is therefore put forward as an additional axiom, the *axiom of non-satiation*.

On the other hand, we are free to choose the origin and the unit scale of utility without affecting preferences. To simplify comparisons, we choose $u(0) = 0$ and a slope of 1 at the origin, that is $u'(0) = 1$.¹⁰

B' , should be the same as our preferences between A and B . But the compounded lotteries A' and B' offer the same two outcomes M and m . To make a choice, according to axiom (A4), we simply have to compare the probabilities of winning the preferred outcome M . These probabilities are $E_A[u] = \sum p_{A;x_i}$ and $E_B[u] = \sum p_{B;x_i}$, that is, renaming as ‘utilities’ the probabilities u_i , they are the expected utilities of lotteries A and B . Therefore the preferred lottery ought to be the lottery with maximum expected utility.

⁹ We ignore pathological cases where, because of crude modelling of outcomes, an infinitesimal increase in cash could apparently lead to vastly different consequences such as having just enough money to get bail or to buy a new house.

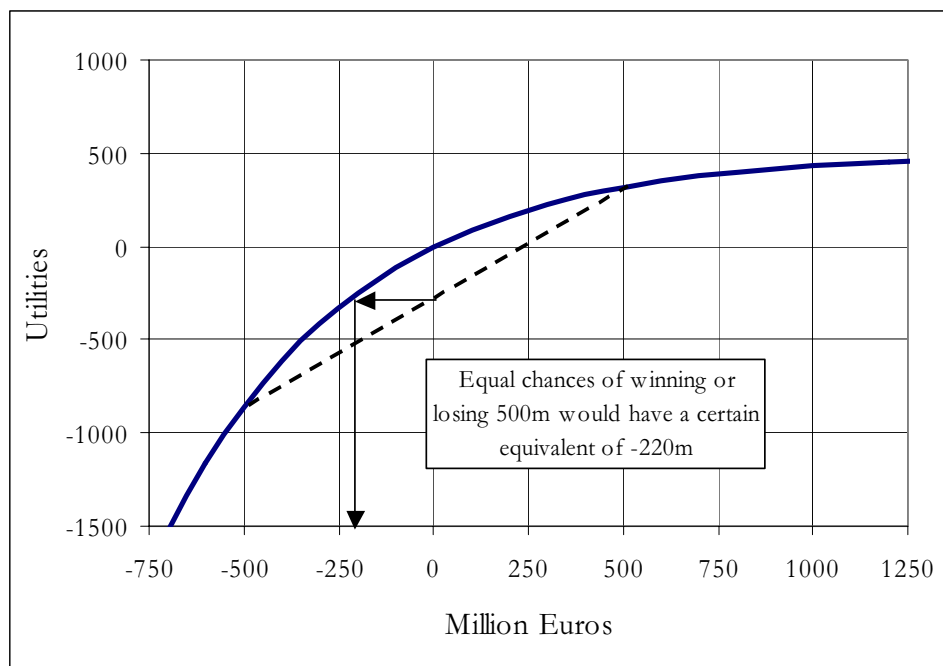
¹⁰ . The expectation operator is linear, that is, $E[(a.u(X) + b)] = a.E[u(X)] + b$, with X a lottery and a and b two scalar constants. Therefore the order of preference set by the maximum EU principle remains unchanged under a positive linear transformation ($a > 0$) of the utility function. Without loss of generality, one may choose a utility scale as in Figure I.A.1.1 where $u(0) = 0$ and the first derivative $u'(0) = 1$, so that for infinitesimal variations around the origin utility and cash have the same unit.

It is also common practice to choose the current level of wealth as the origin of the cash scale so we have zero utility for our current level of wealth. In this case, future wealth is valued against the current level of wealth rather than in absolute terms. We follow this practice here. But we should remember that the level of wealth is unlikely to remain unchanged over time, and this may affect risk attitude. This is not a major drawback as risk attitude may evolve over time anyway and it is therefore prudent to check regularly whether the utility function being used is still representative of risk preferences.

I.A.1.4.3 Risk Aversion (and Risk Tolerance)

It is the curvature of the utility function that captures the risk attitude of a decision maker or a firm. A downward curvature (*concave* utility function) expresses *risk aversion*: the minimum selling price of a risky opportunity is less than its expected value.

Figure I.A.1.1: Describing risk attitude with a utility function



Example I.A.1.1:

Faced with the prospect of winning or losing €500m with equal probabilities, a firm using the utility function plotted in Figure I.A.1.1 would perceive an expected utility of -270 , the average of the utilities of the two outcomes read of the curve: $u(500) = 280$ and $u(-500) = -820$. Now reading back from the curve (black arrows) we find that -270 is the utility of a sure loss of €220m. In other words, the firm would be willing to pay up to €220m to have the risky prospect taken away.

If the curvature were upwards (*convex* utility function), a risky opportunity would be perceived as having greater expected utility than its expected value, which would reveal a *risk-seeking* attitude. Finally, no curvature, that is, a straight-line utility function, would reflect a *risk-neutral* attitude – the expected value of outcomes is the choice criterion. Risk aversion is the norm, at least for business decisions, whereas a risk-seeking attitude is usually regarded as pathological.¹¹ We shall argue later why a utility function should be very smooth (continuous first- and second-order derivatives) and concave for business decisions.

Mathematically, the *curvature* of a twice differentiable function is defined as the ratio of its second-order derivative to its first-order derivative. For a concave utility function such as that in Figure I.A.1.1 the curvature is negative because $u''(x) < 0$. We call minus the curvature *the local coefficient of risk aversion* at x . That is:

$$\text{Local Coefficient of Risk Aversion} = -u''(x) / u'(x).$$

Its inverse is called, quite naturally, the *local coefficient of risk tolerance* at x ; it is expressed in the same monetary units as x and therefore may be easier to interpret.¹² According to the age-old principle of assigning a Greek letter to an unknown parameter, we shall call λ the local coefficient of risk tolerance. Thus

$$\lambda = -u'(x) / u''(x). \tag{I.A.1.1}$$

Stipulating the coefficient of risk tolerance (or the coefficient of risk aversion) over various levels of wealth is equivalent to stipulating a utility function (see Pratt, 1964).

I.A.1.4.4 Certain Equivalence

We call the *certain equivalent (CE)* of a gamble the sure quantity that we would be willing to exchange for the gamble, (i.e. $u(\text{CE}(X)) = E[u(X)]$). In the previous example, minus €220m is the certain equivalent of the project. Clearly, choosing the alternative with the maximum EU is equivalent to choosing the alternative with the maximum CE.

I.A.1.4.5 Summary

Financial risks are gambles. For our purposes, a gamble is a set of *cash-value* outcomes, with some probabilities attached to each outcome. Then, *rational* decisions between financial risks are achieved by:

¹¹ Gambling has always fascinated men. It is not only the subject of gripping stories (such as Dostoyevsky's *The Gambler*) but it also arouses principled and even religious reactions, usually in the form of condemnations. But that is not to say that rational people should necessarily be risk-averse.

- i. defining a utility function $u(x)$, a monotonically increasing function of cash value x ;
- ii. calculating the expected utility $E[u(X)]$ of each gamble X ;
- iii. choosing the gamble that has the maximum expected utility or, equivalently, choosing the gamble with maximum certain equivalent.

Although the current level of wealth is usually taken as the origin of the scale on which future outcomes are valued, each new course of action should not be considered independently of the status quo. The uncertainties we have in the future will depend on what we do today. Each future choice should therefore be considered in the context of *current* uncertainties.

I.A.1.5 Encoding a Utility Function

I.A.1.5.1 For an Individual

The first step in implementing utility theory is to draw a utility function over possible states of wealth of an individual or a firm. It is a tricky exercise best conducted by an experienced and independent experimentalist.

An individual's risk attitude can, in theory, be inferred from a series of decisions, provided the other elements of the decisions (i.e. the outcomes, probabilities, alternatives) are clearly understood by all. It is best, of course, if the problems submitted for decision are:

- i. *Realistic.* One should avoid game playing with all the distortions it may create (e.g. displays of bravado).
- ii. *Meaningful.* The range of monetary outcomes should be on a scale of gains and losses for which we can define a utility function.
- iii. *Clear and simple.* One should avoid ambiguities, or inducements that could lead to misinterpretations of the problem, or biases. In particular, probabilities should be clearly stated and these probabilities should not be so extreme that they cannot be comprehended.

We think of the 'decision maker' as a bank executive or a successful trader. We start by defining a *monetary range of interest* for our decision maker by choosing a minimum and a maximum cash amount, say minus €3 million and plus €10 million. This range should cover the personal impact of decisions she may have to face, for example, insuring her life, deciding whether to accept a new incentive scheme or even, perhaps, whether to commit her company to a new deal such as selling a €20 million credit swap, since the outcome could affect her future.

¹² Mathematicians usually prefer to use the coefficient of risk aversion whereas practitioners usually prefer to use its inverse, the coefficient of risk tolerance; which coefficient is used does not really matter. We shall side here with the practitioners

But we are not going to use the extremes of the range as a starting point for questioning our subject. To begin, we centre the questions around more familiar values. We could ask first:

Question 1: *If you were offered a once-in-a-lifetime opportunity to win x euros or to lose $x/2$ euros with equal chances, for what value of x would you hesitate between taking the gamble and letting the opportunity go by?*

Note that most people consider this gamble attractive for small values of x but if x is very large, the risk becomes a deterrent. The answer will require a good deal of thought and, admittedly, may not be very precise. All we seek at first is an approximate value. The subject may be encouraged to think about realistic situations where she would face a similar type of gamble. But it should be very clear that in the present circumstances the respondent has absolutely no power or responsibility in determining the event of winning or losing (she is not a wizard) nor any means of guessing the outcome correctly (she is not a clairvoyant).

Suppose that after much soul-searching the subject feels that her indifference point is at around $x = €500,000$. Now without loss of generality assign a utility of 0 to a zero gain and a utility of 1 to a gain of €500,000. The answer to the first question should be interpreted as assigning a utility of -1 to a loss of €250,000 since, by equating expected utilities, we must have:

$$u(0) = \frac{1}{2} u(€500,000) + \frac{1}{2} u(-€250,000)$$

or

$$0 = \frac{1}{2} (1 + u(-€250,000)),$$

that is,

$$u(-€250,000) = -1.$$

We would continue by asking for the CE to a few simple gambles and, eventually, push towards the extremes with questions such as the following:

Question 2: *If you were offered (i) a lottery ticket to win €10 million with some probability p , and nothing otherwise, or (ii) a sure prize of €500,000, for which probability p would you be indifferent between taking the lottery or settling for the sure prize?*

Question 3: *If you were asked to pay a €250,000 insurance premium to insure against a potential €3 million loss, what would be the minimum probability of loss that would justify this premium?*

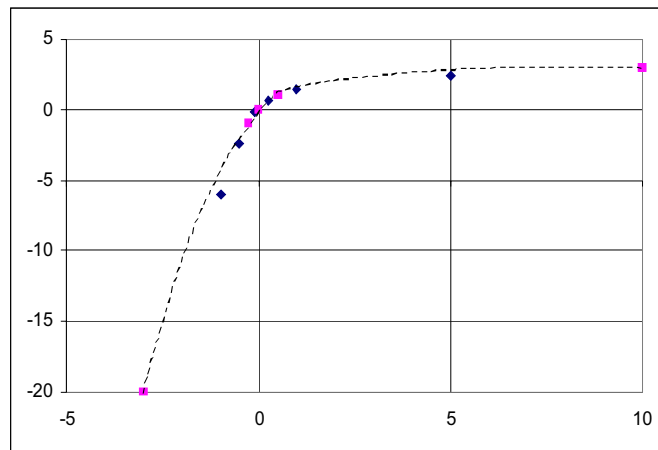
Suppose the answers to questions 2 and 3 are probabilities of one-third and 5%, respectively. Then the utilities of €10 million and minus €3 million can be derived. We should have from:

Q2: $u(€500,000) = (1 - p) u(0) + p u(€10m)$ or $1 = (1/3)u(€10m)$ hence $u(€10m) = 3$

Q3: $u(-€250,000) = (1 - p) u(0) + p u(-€3m)$ or $-1 = 0.05u(-€3m)$ hence $u(-€3m) = -20$.

Putting these results together gives five points for the subject’s utility curve. These are marked with squares in Figure I.A.1.2. We imagine that other similar questions produced the results marked with diamonds. Now we can draw a tentative, freehand smooth curve through these points and show the result to the trader. It is a useful basis for further discussions. Perhaps some additional questions should be asked to confirm some of the previous answers, perhaps some features of the curve will be perceived as anomalies, or perhaps the trader will want to impose some properties on the curve. For a comprehensive review of utility elicitation techniques, see Farquhar (1984); see also Wakker and Deneffe (1996) for a trade-off method designed to avoid probability distortions in encoding.

Figure I.A.1.2: Encoding of a personal utility function



I.A.1.5.2 For a Firm

Encoding a utility function for a firm presents a much bigger challenge than for an individual. Two questions arise immediately:

Should there be a single risk attitude for all decisions in the firm?

Whose risk attitude should be encoded?

In favour of a single utility curve for the whole firm we have the traditional argument of consistency. Arguably, a key benefit of formalising risk attitude is to be able to achieve harmony between business units. Seen from the outside, it would not make sense if one unit owned a particular project that was perceived as too risky whilst another unit would find the same project attractive. A wily outsider could ask to be paid by the first unit to take the project away and then

go to the second unit to sell them the same project.¹³ It is for similar reasons that firms try to use a single discounting factor for evaluating the net present values of various projects.

But consistency must be balanced against practicality. In a complex organisation, decisions are taken at all levels, from the most routine operating decisions at low departmental levels to the most strategic decisions taken at board level. For the organisation to function efficiently each decision analysed at an intermediate level must rely not only on a framework of higher-level strategies but also on general assumptions about lower-level decisions and their likely consequences. A uniform risk attitude could mean that relatively more risks will be taken at lower levels, with the consequence that it becomes more difficult to plan. And the cost of disruption, of not being able to stick to a plan, may be high although difficult to analyse. Rather than to develop complex contingency plans and rely on cross subsidies between units (which might de-motivate managers and staff), it may appear more expedient to limit risk taking at lower levels.

This leads to the second problem: whose risk attitude should we encode? To paraphrase a well-known and perhaps misunderstood quote, there is no such thing as a firm.¹⁴ There is only a group of individuals (employees managers, stakeholders and clients) who share some common interests. But they have no reason to share the same risk attitude. Shareholders may be perfectly happy to see the firm take a large gamble, especially if the firm is already in a distressed situation. They own out-of-the-money call options on the assets of the firm. The share price is likely to respond very positively to the undertaking of a profitable although risky venture. Obviously the bondholders may totally disapprove of this venture. Employees may share an intermediate view, whilst managers are supposed to act in the best interests of all parties – including themselves, naturally.

It may be, however, that this heterogeneity of interests would push management to negotiate a common risk-attitude objective and have it agreed by all parties. We observe that, in the financial industry, regulators are setting minimum capital requirements relative to risks undertaken in order to protect the wider interests of creditors and the public in general. But many financial firms have gone one step further, trying to define a desirable (if not optimal) level of capital consistent with the risks undertaken and the objectives of their stakeholders.

It would seem reasonable, therefore, that a firm's risk attitude to strategic decisions should be agreed at board level. Then the risk policy could be delegated to divisional and departmental

¹³ I must confess to having done exactly that with two departments of the same investment bank.

levels, leaving some freedom at each level of management to increase the degree of risk aversion that should be used at lower levels of decision making. A balance has to be struck between too much risk aversion at lower levels that would kill interesting opportunities and too much risk tolerance that would be disruptive to the planning process.

A great advantage of an explicit risk policy is that the personal interests and risk attitudes of individuals are less likely to interfere with the common good of the firm's stakeholders as interpreted by the board. It is not only inevitable but also desirable that the interests of individual managers be taken into account, but it is very difficult to design incentive schemes that will align individual interests with the common good. By having an explicit risk policy, individuals have a chance to justify their decisions to do what is best for the firm rather than be judged only on the results of their decisions, which to some extent are left to chance, and do what is best to promote and protect exclusively their personal interests. Perhaps for this reason, it is not unusual to find that, after thorough discussions, the consensus opinion on the appropriate risk attitude a firm should adopt is much more risk-tolerant than the average risk attitude of the executives interviewed. Such findings, as well as a detailed procedure for eliciting a corporate risk policy, are reported by Spetzler (1968).

I.A.1.5.3 Ironing out Anomalies

A first sketch of an empirically derived utility function often reveals features that, on closer inspection, the owner may not agree with. A mathematician can deal with any utility function, but a decision maker must verify that she is happy with the quantification of her risk attitude.

Two common features that perhaps would on second thoughts be undesirable are:

- (i) a 'kink' around the current level of wealth ($x = 0$);
- (ii) a reversal of curvature from negative with positive outcomes to positive with negative outcomes.

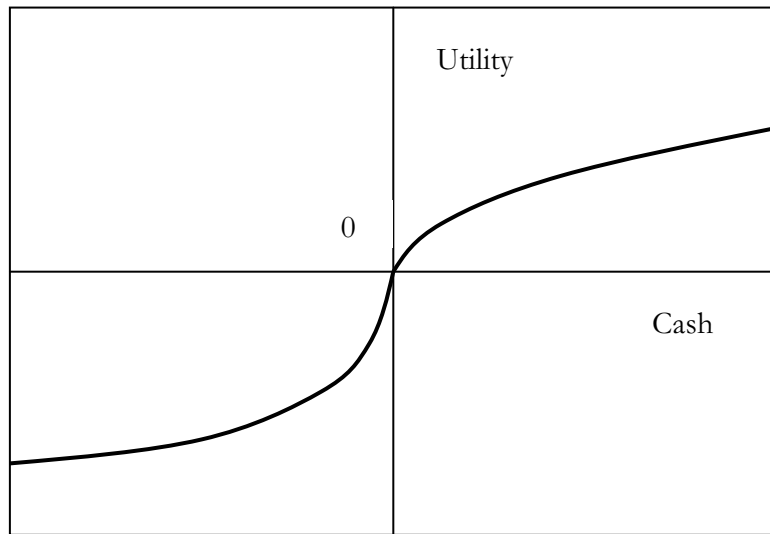
Figure I.A.1.3 illustrates these two features. The kink around the origin indicates a high level of risk aversion for any change from the status quo. This supports the paradoxical view jokingly attributed to old bankers: 'Change, even for the better, is always upsetting'.¹⁵ But there are two main arguments against the kink at the origin, at least in a business context. The first is that the current economic value of a business is uncertain, even somewhat intangible, as one should try to

¹⁴ The original quote by Mrs. Thatcher was: 'There is no such thing as society' (interview for *Women's Own* magazine, 23 September 1987).

¹⁵ Kahneman and Tversky noted this pervasive attitude: people seem more sensitive to deviations than to absolute levels; it is not compatible with a von Neumann–Morgenstern utility based on wealth. Kahneman and Tversky propose instead a 'prospect theory' with choices based on the optimisation of a value function $v(x, m)$, where m stands

factor in the value of future prospects,¹⁶ hence any new choice with limited consequence has only a mild impact on the current state of uncertainty. This should smooth the utility curve around the origin as we do not exactly know where the origin is. The second argument is that, even in the absence of uncertainty about the future, things change over time. So, particularly in business, one should not be too attached to the present. The kink at the origin of the utility curve may be no more than the illusion of a no-risk, stable status quo.

Figure I.A.1.3: A utility function with typical ‘anomalies’



The reversal in curvature from a risk-averse attitude towards gains (cash > 0) to a risk-seeking attitude towards losses (cash < 0) has been noted before. It is not always present (for instance, most people buy insurance even when they are not forced to) and it may be less prevalent among firms than it is among individuals. One explanation is that the consequences of large losses may not be adequately described. It would be naïve to pretend that an individual could lose anything close to her net worth and would be left to starve; in our civilised societies there are bankruptcy laws and other safety nets before anyone reaches this point. Likewise, limited companies cannot lose more than their capital and extreme risks are borne by others than company executives.

When the consequences of large losses are properly described and there is still a wealth level (or indeed several) where risk attitude flips between aversion and risk-seeking, consider the situation where wealth fluctuates around one of these critical levels. If one day the net wealth is in the risk-seeking region, the firm would be willing to pay a fee to play a fair game (a game with zero expected value); if the next day the net worth has increased and stands in the risk-averse region,

for current wealth and x for wealth increments. They go on to show that, empirically, the value function is only mildly dependent on m but markedly dependent on x . That is, it is one's *change* in wealth that matters!

¹⁶ Behaviourists have also shown that we tend to underestimate future risks. In business circumstances, especially, we tend to be overconfident.

the firm would be willing to pay an insurance premium to avoid taking the same fair game. In other words, as wealth fluctuates around the critical level, the firm could be used as a money pump, a less than attractive prospect.

In conclusion, taking into account the uncertainty about the net worth of a firm and its fluctuating nature, the risk preference of a firm should be expressed by a smooth utility function with a *single* type of curvature. And that curvature will typically be *negative* as providers of funds (both equity and debt) are typically risk-averse.

I.A.1.6 The Mean–Variance Criterion

Armed with a utility function, one may evaluate the certain equivalent of any project whose outcomes fall within the domain over which the utility function has been defined. Utility functions are needed to evaluate decisions where low probability but extreme outcomes may occur, for example, a decision to insure or not a large risk at a certain premium or a decision to make a strategic investment. But smaller, more frequently encountered risks may be evaluated by using a simplified expression of risk attitude.

I.A.1.6.1 The Criterion

As we noted in Section I.A.1.4, it is the *curvature* of the utility function that expresses the level of risk aversion of a decision maker. We make use of this property to state a simplified version of the maximum EU principle, which is called the *mean–variance criterion*.¹⁷

Consider a ‘gamble’ X with expected value $E[X]$ and variance $\text{Var}(X)$. For instance, the gamble may represent a portfolio return. The mean–variance criterion for choosing among mutually exclusive gambles is:

$$\text{maximise } \{E[X] - \text{Var}(X)/2\lambda\}, \quad (\text{I.A.1.2})$$

where λ characterises the risk tolerance of the decision maker. Note that, if the outcomes are dependent on other uncertainties the decision maker already faces, call them Y , the variance

¹⁷ The following derivation of the mean–variance criterion is not examinable in the PRM. With any well-behaved utility function, when the range of possible outcomes of a gamble is small relative to the risk tolerance of the decision maker, the certain equivalent of the gamble can be approximated by making use of the local curvature of the utility function. Consider a gamble X with expected value $E[X]$ and variance $\text{Var}(X)$, and a utility curve $u(x)$. From the first two terms of a Taylor series expansion (see Section II.C.3.3) of $u(x)$ around the expected value $E[X]$ we obtain: $E[u(x)] = u(E[X]) + (1/2) u''(E[X])\text{Var}(X)$. We equate this expected utility to the utility of the certain equivalent $\text{CE} = E[X] - D$, where D is a risk discount. Again, using a Taylor series expansion, $u(E[X] - D)$ can be approximated by $u(E[X]) - u'(E[X])D$; hence the first-order approximation for the risk discount is $D = \frac{1}{2}(u''/u') \text{Var}(X)$; therefore,

$$\text{CE}(X) \approx E[X] - \frac{1}{2} (u''/u') \cdot \text{Var}(X).$$

The maximum EU (or maximum CE) principle reduces to a linear function of mean and variance in which the relative weight given the variance is half the local curvature of the utility function.

estimates in equation (I.A.1.2) should be the marginal contribution of the new gambles to the total variance. Then the mean–variance criterion becomes:

$$\text{maximise } \{E[X] + E[Y] - (\text{Var}(X) + 2 \text{Cov}(X, Y))/2\lambda \}. \quad (\text{I.A.1.3})$$

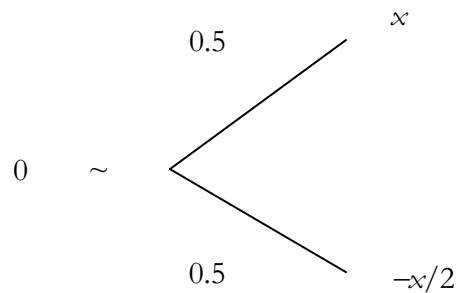
For instance, if one already holds a portfolio with returns Y and is considering investing in an additional portfolio with returns X , the decision will be affected by the diversification potential, that is, the covariance between Y and X (see Chapter I.A.3).

I.A.1.6.2 Estimating Risk Tolerance

Interestingly, the mean–variance criterion can be used to obtain a quick estimate of the risk tolerance of a firm or individual. Consider an opportunity to gain x or to lose $x/2$ with equal chances independently of any other risks the firm already incurs. This is depicted in Figure I.A.1.4.

Figure I.A.1.4: Quick assessment of the risk tolerance coefficient

The coefficient of risk tolerance is approximately the value of the pay-off x for which the decision maker is indifferent between taking the following gamble or abandoning it



For this gamble X ,

$$E(X) = \frac{1}{2} x + \frac{1}{2}(-x/2) = x/4$$

and

$$\text{Var}(X) = \frac{1}{2}(x - x/4)^2 + \frac{1}{2}(-x/2 - x/4)^2 = 9x^2/16.$$

If x is very small, the variance is negligible compared to the expected value, expression (I.A.1.2) is positive and therefore the gamble is worth taking (with any well-behaved utility function, it is worth taking a small stake in any ‘good’ gamble, that is, any gamble with a positive expected value). If we now increase x , the variance increases faster than the expected value and at some point expression (I.A.1.2) becomes negative; the gamble is no longer worth taking. In fact, expression (I.A.1.2) is zero when $x = 8\lambda/9$. Hence we set $\lambda = 9x/8$, where x is the value for which she is indifferent between taking the gamble and not.

Example I.A.1.2:

If the point of indifference of an individual is reached when $x = €100,000$ then that is approximately her coefficient of risk tolerance. More precisely $\lambda = 9x/8 = €112,500$, although such a degree of accuracy is rather dubious.

I.A.1.6.3 Applications of the Mean–Variance Criterion

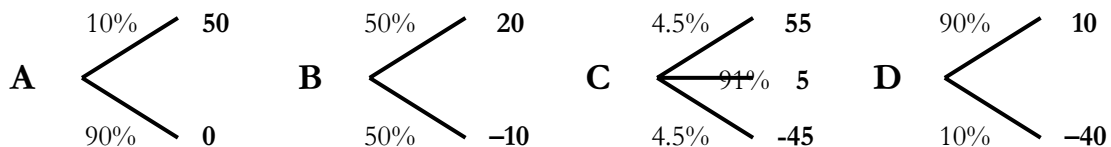
The mean–variance criterion and, more generally, mean–variance analysis¹⁸ have been used in a multitude of areas in finance and economics: portfolio selection, capital budgeting, optimal hedging, optimal consumption and insurance decisions. As examples of early discussions and applications, see Markowitz (1959) and Sharpe (1964). Also Section I.A.3.2 describes the application of mean–variance analysis to portfolio optimisation.

The mean–variance criterion is a useful simplification of the principle of maximum EU when the range of outcomes under consideration does not exceed plus or minus one coefficient of risk tolerance. It is also a good approximation when analysing bigger gambles with special classes of utility functions and/or distributions of outcomes (more details are given in Appendix I.A.1.B). Using the mean–variance criterion rather than the maximum EU principle simplifies a number of optimisation problems where risks must be balanced against returns. Only the first two moments of pay-off distributions need be known and mathematical techniques such as quadratic programming may be used.

However, some academics as well as practitioners are not satisfied with this criterion. They point to circumstances where choices guided by mean–variance considerations might deviate markedly from choices guided by the principle of maximum EU. They have also tried to derive better approximations, some based on more complex, non-linear functions of expectation and variance – these are referred to in the literature as expectation–variance or ‘EV’ criteria – some based on other measures of rewards and risks. We discuss some of these other criteria in the next section.

Figure I.A.1.5: Four gambles with equal expected values and variances

Do you feel the following gambles are all equally attractive?



To judge whether you personally might not be satisfied with any EV criterion, consider the four gambles in Figure I.A.1.5. The probabilities of the outcomes are expressed as percentages. Think of the unit of payment as being large enough (thousands of euros?) so that gains or losses represent a few percentage points of your net worth. Which gamble would you prefer to take? Gamble A will appear attractive (nothing to lose) but gamble B and especially C and D may appear too risky. If that is the case for you, then you will not be satisfied with an EV criterion! That is because all four gambles have the same expected value of 5 and the same standard deviation of 15.

I.A.1.7 Risk-Adjusted Performance Measures

Maximising expected utility will always tell us which of several mutually exclusive gambles (e.g. which portfolio) an investor should prefer. We now consider a specific problem known as the *static portfolio selection* problem. We start with a set of risky assets that can be bought or sold at the same price in any quantity to create portfolios (linear combinations of assets). Asset returns – gains or losses per unit investment over a given time – are described by a joint probability distribution. An investor has some capital to invest; how should he allocate it? If we apply the maximum EU principle, we know that the optimal investment will generally depend on the investor's risk attitude.

But there are circumstances in which *all* risk-averse investors would agree that some portfolios of risky assets are worse than others. They are left to choose their preferred portfolio from a family of better portfolios – the *efficient portfolios* – according to their personal risk attitude. In special cases defined as market equilibrium conditions, all investors will even agree on the same optimal mix of risky assets, which we call the *market portfolio*. Their personal risk attitude will only determine the amount to be invested in the market portfolio and the amount to be invested in or borrowed against the risk-free asset. Full details are given in Chapter I.A.3.

In such circumstances it may be possible to choose a risk metric that is proportional to the quantity invested in the market portfolio. For instance, the standard deviation of portfolio returns is such a risk metric. Then, on an *expected return* versus *risk* diagram such as Figure I.A.3.9 all combinations of the market portfolio and the risk-free asset will lie on a straight line. All efficient portfolios will have the same ratio of expected excess returns (above the risk-free rate) over risk.

¹⁸ Mean–variance analysis refers to any risk–reward analysis based only on mean and variance but not necessarily applying the mean–variance criterion (see Section I.A.3.2).

This has given rise to the concept of *risk-adjusted performance measure* (RAPM). These measures, usually ratios of expected returns to risks, are designed to provide a preference ranking of risky opportunities acceptable to a majority of investors, assuming only that these investors are risk averse in a simple sense.

RAPMs have become extremely popular because they are easy to calculate and, oddly, do not seem to require a statement of personal risk attitude. In fact, their range of applicability is very limited: asking which investment is best regardless of an investor's risk preferences is about as meaningful as asking which music composer is best regardless of a listener's musical inclinations. For this reason we have a large (and still growing) number of different definitions for RAPMs. Each different RAPM induces a different preference order, and we have endless debates as to which is the best! In this section we review the traditional RAPMs, including the Sharpe ratio, the Treynor ratio, the Jensen alpha, RAROC and RoVaR, and some more recent ones, the kappa indices (Sortino) and omega index. In each case, we describe the circumstances in which the RAPM could be applied and the link, if any, with utility theory.

I.A.1.7.1 The Sharpe Ratio (see also Sections I.A.3.7 and I.A.4.4.1)

We add two conditions to the static portfolio selection problem described above:

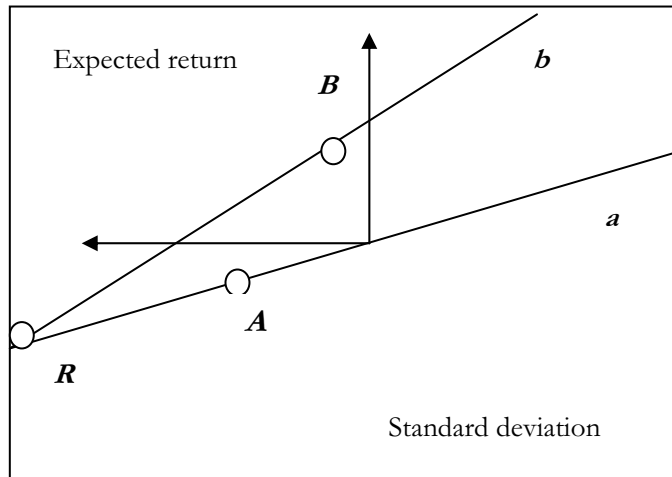
- i. There is a *risk-free* asset returning the risk-free rate r . It is possible to buy or sell the risk-free asset in any amount – in other words, it is possible to borrow as well as invest any amount at the risk-free rate.
- ii. Investors' risk preferences can be described in terms of expected value and standard deviation only; they are *risk-averse* in the minimal sense that for any given level of risk, investors prefer the portfolio yielding maximum expected return (or, equivalently, for any level of expected return they prefer the portfolio yielding minimum risk).

Sharpe (1964) demonstrated that under these conditions optimal portfolios maximise the ratio

$$SR = (\mu - r)/\sigma, \quad (\text{I.A.1.4})$$

where μ represents the expected return of a portfolio, $(\mu - r)$ its excess return over the risk-free rate and σ the standard deviation of excess return. *SR* has become known as the *Sharpe ratio*.

Figure I.A.1.6: Choice of either of two risky asset



The reasoning is straightforward: position two risky and mutually exclusive portfolios A and B on an expected return versus standard deviation diagram. Also position the risk-free asset R as in Figure I.A.1.6 and draw the straight lines RAa and RBb . Any point along these lines represent a portfolio that is attainable by combining long positions in either A or B respectively and positions – either long or short – in R .¹⁹ In our illustration, for any portfolio on the RAa line, there are portfolios on the RBb line that have both less risk and larger expected return and which therefore ought to be preferred to the portfolio on the RAa line. This is true as long as the slope of line RBb is greater than the slope of line RAa , that is, as long as the Sharpe ratio of portfolio B is greater than the Sharpe ratio of portfolio A .

The Sharpe ratio can also be deduced from a special application of the mean–variance criterion. Suppose we finance the purchase of a quantity q of a risky portfolio (μ, σ) at the risk-free rate r . The new risk is independent of other risks we may have. According to the mean–variance criterion (I.A.1.3), the marginal CE contribution of the investment is:

$$CE(q) = q(\mu - r) - (q^2 \sigma^2 / 2\lambda) \tag{I.A.1.5}$$

The investment quantity q that maximises the CE is $q = \lambda(\mu - r) / \sigma^2$ and the corresponding maximum CE is $CE^* = \frac{1}{2} \lambda((\mu - r) / \sigma)^2$. Thus the portfolios that contribute to the maximum increase in CE^* are those that maximise SR , whatever the level of risk tolerance of the investor.

The SR rule is justified *only if conditions (i) and (ii) above are fulfilled and the investment maximising the expected utility contribution is contemplated*. This is a very restrictive set of conditions. Otherwise, for

¹⁹ The expected value of a portfolio is a weighted average of the expected value of its components. The standard deviation is proportional to the amount invested in the risky portfolio.

instance when risk aversion cannot be expressed in terms of expected value and standard deviation alone, the SR rule may lead to paradoxical conclusions. Here is a simple illustration.

Example I.A.1.3:

The two mutually exclusive investments A and B depicted in Figure I.A.1.7 have the same characteristics except that the best outcome with B is greater than the best outcome with A . Stochastic dominance indicates that B should be preferred to A . But that is not what the Sharpe ratio rule indicates. The expected excess returns of A and B are 10% and 14% respectively, and the standard deviations of excess returns are 14.14% and 23.83% respectively. Therefore $SR(A) = 0.707$ and $SR(B) = 0.587$. The Sharpe ratios indicate that A should be preferred to B , in flagrant violation of stochastic dominance – a not so sharp answer!

Figure I.A.1.7: Choice of two mutually exclusive investments



I.A.1.7.2 RAPMs in an Equilibrium Market

If new investments introduce risks that are related to existing risks, the Sharpe ratio, like the mean–variance criterion, should be calculated with reference to the marginal risk contributions of the new investments. Also, when new investment opportunities are not exclusive and are correlated, the Sharpe ratios of all possible combinations of new investments should be calculated to decide on the best combinations. For example, it would be quite possible that between two new investment opportunities the first would have a higher Sharpe ratio than the second when estimated separately, yet the second could be preferred to the first because, unlike the first, it is negatively correlated with existing risks and consequently makes a lower marginal contribution to total risk.

I.A.1.7.2.1 The Treynor Ratio and Jensen’s Alpha

Investments in traded securities are usually considered in the context of a securities market in equilibrium with a large number of participants and alternative investments, including a risk-free asset. The market as a whole can then be considered as the optimal market portfolio. To analyse new investments in a market context or to assess the performance of portfolio managers, Treynor (1965) and Jensen (1969) proposed alternatives to the Sharpe ratio. The *Treynor ratio* rule is to choose the portfolio that maximises $TR = \alpha/\beta$, where α and β are the parameters of a linear

regression of the excess return of a portfolio against the average excess return of the market portfolio:²⁰

$$\mu - r = \alpha + \beta(\mu_m - r). \quad (\text{I.A.1.6})$$

In the regression $\mu_m - r$ is the market excess return; $\beta = \rho\sigma/\sigma_m$ is the sensitivity of the portfolio excess return to changes in the market portfolio excess return, ρ , σ and σ_m being respectively the correlation between the portfolio and market returns, the standard deviation of the portfolio return and the standard deviation of the market return; and α is the portfolio excess return not explained by the market excess return.

The Treynor ratio has an intuitive feel: the greater the value of α and the smaller the value of β , the larger the expected return and the lower the correlation of the portfolio with the market, two desirable features. Jensen, on the other hand, suggests that α alone should be viewed as a measure of true performance of a portfolio and consequently of the skills of the portfolio manager (see Section I.A.4.4.2). Both the Treynor and Jensen criteria seem more adapted to an *ex post* performance evaluation than to an *ex ante* investment analysis.

I.A.1.7.2.2 Application of the Treynor Ratio

If we apply the mean–variance criterion (I.A.1.3) to the static portfolio selection problem (with the assumption that all risks that are not correlated with the market can be well diversified and are therefore negligible) we come to the following conclusion:

By maximising their certain equivalents the actions of all investors, irrespective of their risk aversion, will bring asset prices to an equilibrium. In this equilibrium we have, for any risky asset

$$SR = \rho SR_m.$$

This linear relationship between the asset Sharpe ratio and the Sharpe ratio of the market portfolio is fundamental to the *capital asset pricing model* (CAPM); it is fully discussed in Chapter I.A.4. The CAPM is the same linear relationship as the regression equation (I.A.1.6) but with $\alpha = 0$. According to the CAPM, on an expected return versus beta plot, all assets and portfolios should lie on the same straight line running through the risk-free asset and the market portfolio; this line is called the *security market line* (see Section I.A.4.3). If we have reasons to believe that an asset or portfolio lies above the security market line, this asset is underpriced. It is therefore a good buy opportunity.

²⁰ For convenience, a broad market index is often considered representative of the market portfolio.

How can we rank investment opportunities that we think would lie above or below the security line? We apply the mean–variance criterion and consider only the *systematic risks* (i.e. the risks that are correlated with the market) as significant risks. Specific risks are diversifiable (see Section I.A.4.2.2). This leads to the same equation as (I.A.1.6) but with the total risk σ replaced by the systematic risk $\rho\sigma$. We conclude that in a securities market, we should invest the optimal amounts in portfolios that maximise $(\mu - r)/\rho\sigma$, the excess return per unit of systematic risk. But that is equivalent to maximising α/β , the Treynor ratio.²¹

I.A.1.7.2.3 Application of Jensen's Alpha

The Jensen alpha is the answer to a slightly different problem: if, in a market environment, one seeks to invest a small amount in a new asset or portfolio, which should it be? The answer is to invest in the portfolio with the largest Jensen alpha. In the same way, in a universe of independent assets the answer would be to prefer small investments in the assets promising the largest expected returns, irrespective of risks.

Why, then, is alpha widely used by fund managers? Perhaps because only small investments are considered; perhaps also because the Jensen alpha generalises easily to multifactor market models. Securities markets do not behave exactly as predicted by CAPM. *Arbitrage pricing theory*, as developed by Ross in 1976, proposes an extension to multifactor risks. In this theory alpha remains a scalar even when multiple betas are introduced, one for each risk factor. It is more difficult to generalise the Treynor ratio to multifactor risks than the Jensen alpha, although it is not infeasible.²²

I.A.1.7.3 Generalising Sharpe Ratios

The violation of stochastic dominance shown in some circumstances by the Sharpe ratio rule (or, equivalently, by the mean–variance criterion) is more than a mere curiosity. Often the risks to be compared are not similarly distributed. In finance, the dynamics of asset prices are often modelled with a constant-volatility geometric Brownian process implying that log-returns are normally distributed (see Section I.A.8.7). But there is ample empirical evidence to suggest that volatility is not constant and is difficult to forecast, that log-returns have heavier tails than normal distributions (positive excess kurtosis) and are often asymmetrical (non-zero skew) – see Sections II.B.5.5 and II.B.5.6. Even if basic assets were similarly distributed, many portfolios contain

²¹ The certain equivalent of the investment is maximised when $(\mu - r)/\rho\sigma$ is maximised, or, substituting $\alpha + \beta(\mu_m - r)$ for $(\mu - r)$, when $\alpha/\rho\sigma + \beta R_m$ is maximised, which amounts to maximising α/β . Note the special case when $\beta = 0$: the Treynor ratio becomes infinite, indicating that any investment not correlated to the market but yielding an expected return above the risk-free rate is infinitely desirable. Obviously, if one began to invest heavily in that opportunity, its specific risk (risk not correlated with the market) would become significant and impossible to diversify; the diversification assumption supporting the Treynor ratio would break down.

²² See Huebner (2003).

instruments such as options that are non-linear in the basic assets or are actively managed so as to create skewed and heavy-tailed returns. Most securities are also vulnerable to default of their issuer, creating a small probability of a large loss, that is, a negative skew. Finally, many risks are not obviously linked to market returns; there are many business and operational risks with heterogeneous distributions.

Academics and practitioners have therefore tried to design choice criteria that would be simpler to implement than maximising expected utility (this requires a utility function and full probability distributions) and yet would have a wider range of applicability than the mean–variance criterion or its equivalent for ranking investments, the Sharpe ratio.

One avenue of research is to base preferences on simple risk–reward value functions that, on the one hand, are simpler to evaluate than expected utility and, on the other hand, are richer than mean–variance functions in that they would capture other risk characteristics of the choice portfolios. A simple family of risk–reward functions consists of linear combinations of a reward metric, typically the expected value, and a more ‘appropriate’ risk metric than variance. A host of candidate risk metrics have been put forward, some *ad hoc* such as value-at-risk (see Chapters III.A.2, III.B.6 and III.C.3) and some more soundly grounded, such as semi-variance (see Section II.B.5.4). We give a few examples in the next section.

Two arguments, often but not necessarily compatible, have been used to devise an appropriate risk metric. One is that the value function in which it will be used should be compatible with the existence of a utility function (i.e. the value function should lead to the same preference order). The other is that a risk metric should have a few intuitive properties to account for aggregation of risks, diversification, hedging, etc. This second approach has led to the specification of *coherent risk measures* with desirable properties (see Artzner *et al.*, 1999).²³

I.A.1.7.3.1 The Generalised Sharpe Ratio

Compatibility with the existence of a utility function has led to two types of generalisations of the Sharpe ratio. One, pioneered by Hodges (1997), seeks to extend the Sharpe ratio to apply to the

²³ Artzner *et al.* (1999) propose four axioms for the coherence of a risk metric ρ . Given gambles X and Y with cumulative distribution functions F_X and F_Y , we require:

- i. Sub-additivity $\rho(X + Y) \leq \rho(X) + \rho(Y)$;
- ii. Homogeneity $\rho(aX) = a\rho(X)$, for any scalar a ;
- iii. Monotonicity: $\rho(X) \geq \rho(Y)$ if, for any scalar u , $F_X(u) \geq F_Y(u)$ (stochastic dominance);
- iv. Risk-free condition: $\rho(X + b) = \rho(X) - b$, for any scalar b .

The first axiom ensures that the risk of an aggregate portfolio is no greater than the sum of the risks of its constituents, a property that enables risk budgeting. The first two axioms ensure that the risk of a diversified portfolio is no greater than the corresponding weighted average of the risks of the constituents. The third axiom ensures that stochastic dominance is preserved. The last axiom, which we have rewritten in a time-independent context, indicates that the risk measure is defined on a monetary scale and can be offset by cash amounts (e.g. capital).

comparison of any return distributions. Hodges assumes an exponential utility function (see Appendix I.A.1.B) and seeks the optimal quantity q to invest in a risky portfolio financed at the risk-free rate r . If the return of the risky investment is normally distributed with mean μ and variance σ^2 (see Section II.E.4.4), then an investment q produces the same certain equivalent $CE(q)$ as in (I.A.1.5) and therefore the same optimum value of q . The corresponding maximum EU is

$$EU^* = (1/\lambda)[1 - \exp(-1/2((\mu - r)/\sigma)^2)]. \quad (\text{I.A.1.7})$$

Therefore, as we have already found, investors maximising EU should prefer to invest in the portfolio that maximises the Sharpe ratio *whatever* their risk tolerance λ . Since (I.A.1.7) can be rewritten as

$$SR = [-2\ln(1 - \lambda EU^*)]^{1/2}$$

and since utility functions are defined only within a positive linear transformation, Hodges defines a generalised Sharpe ratio (GSR) as:

$$GSR = [-2\ln(-EU^*)]^{1/2}. \quad (\text{I.A.1.8})$$

The GSR is equivalent to the traditional Sharpe ratio for ranking portfolios with normally distributed returns and when the utility function is exponential. But its range of applicability extends to *any type* of return distribution. The drawback, of course, is that it is restricted to exponential utility functions²⁴ and it requires an expected utility maximisation. However, its advantage is that it produces a dimensionless index for ranking risky portfolios of any type in a manner consistent with expected utility and without having to specify a coefficient of risk tolerance. The GSR is also a coherent risk measure according to the axioms of Artzner *et al.* (see footnote 24).

I.A.1.7.3.2 The Adjusted Sharpe Ratio

The second approach to compatibility with utility theory uses a Taylor series expansion of a utility function to account for higher moments of the return distribution than just mean and variance. The derivation of such an *adjusted Sharpe ratio* (ASR) is beyond the scope of the PRM exam and we simply state it here:

$$ASR = SR[1 + (\mu_3/6)SR - ((\mu_4 - 3)/24)SR^2], \quad (\text{I.A.1.9})$$

where μ_3 and μ_4 the third and fourth standardised central moments of the return distribution, that is, the skewness and the kurtosis (see Section II.B.5.5 and II.B.5.6). The ASR exhibits not only

²⁴ Similar generalisations of the Sharpe ratio can be obtained for other classes of utility functions such as the members of the power family described in Appendix I.A.1B.

variance aversion but also aversion to negative skewness ($\mu_3 < 0$) and to positive excess kurtosis ($\mu_4 - 3 > 0$).

Example I.A.1.4:

We calculate the ASR and GSR for the two investments *A* and *B* in Figure I.A.1.7. The results are compared in Table I.A.1.1.

Table I.A.1.1: RAPMs for choosing between two mutually exclusive investments

Comparison of the Sharpe ratio (SR), adjusted Sharpe ratio (ASR) and generalised Sharpe ratio (GSR) for the two risky investments in Figure I.A.1.7

	SR	ASR	GSR
Investment A	0.707	0.685	0.691
Investment B	0.587	0.709	0.718

Both modified Sharpe ratios indicate that investment *B* should be preferred to investment *A*, as logic dictates. Note, however, that ASR is a simple adjustment to the Sharpe ratio and is not guaranteed to be always consistent with maximum expected utility or even with stochastic dominance.

I.A.1.7.4 Downside RAPMs

Psychologically as well as practically, economic agents are often more concerned by the downside risk, or risk of underperformance, than the upside risk, or risk of overperformance. That is true for fund managers who like to compare their performance to a benchmark or a peer group. It is also true for managers and regulators who want to limit the risk of insolvency of firms to very low levels and thus help maintain credit ratings and avoid financial debacles. For instance, investment-grade firms (rated BBB and above) exhibit historical default probabilities of the order of 0.5% over a year, and AA-rated firms probabilities of the order of 0.05% only.

With a concave utility function, the negative utility impact of a loss is greater in absolute terms than the positive utility impact of the opposite gain. And utility theory makes use of the full probability distribution of outcomes. However, some simplified applications of the maximum EU principle, such as the mean–variance criterion leading to the use of the Sharpe ratio, rely on a single, symmetric characteristic of a probability distribution, in this case the variance. Therefore they may not be safe to apply to distributions that are obviously asymmetric. We illustrated this problem in the previous section and suggested an adjustment to the Sharpe ratio to account for skewness and excess kurtosis.

I.A.1.7.4.1 RAROC

Several other approaches have been used to simplify calculations and to circumvent a detailed statement of risk attitude whilst stressing downside risks. Thus, one may impose a downside risk constraint, for example that potential losses should not exceed a given threshold with more than a certain probability. That is what banking regulators have chosen to do.²⁵ Banks are free to take on risks as long as their eligible capital – as defined by the regulator – is sufficient to ensure that their probability of default over a year remains less than a small percentage. The difference between the current value of a business and a specified loss quantile α at a specified time horizon T is called the *value-at-risk* and is more fully described in Part III of the *Handbook*.

As the risk constraint is translated into a capital requirement, it has a cost that must be taken into account. This is what financial institutions do when they develop their own RAPMs. They look at maximising the ratio of net returns over the corresponding capital requirements. Bankers Trust were the first to publicise the use of a home-grown RAPM when they introduced the *risk-adjusted return on capital* (RAROC) in the late 1970s. There is now a whole zoo of such measures with minor definitional variations; Matten (2000, pp. 146–149) gives a guided tour. In this *Handbook* there is an entire chapter devoted to capital allocation and RAPMs. This chapter is so fundamental to risk management that we have placed it as preliminary reading to all chapters in Part III – it is thus numbered Chapter III.0! The generic form of a RAROC is:

$$\text{RAROC} = (\text{Expected return net of costs and expected losses}) / (\text{Economic capital}), \quad (\text{I.A.1.10})$$

where the numerator may include an adjustment for risks. For example, the cost of funding may be related to risk. The *economic capital* is an estimate (internal to the firm) of the amount of capital necessary to cover possible losses up to a certain confidence level chosen by the firm, such as 99.9%, per year. More details are given in Section III.0.2.

A downside RAPM this is closely related to RAROC is the *return over VaR* (RoVaR), defined as

$$\text{RoVaR} = (\text{Expected return net of costs and expected losses}) / (\text{VaR}), \quad (\text{I.A.1.11})$$

The difference between (I.A.1.10) and (I.A.1.11) is that whilst economic capital is normally assessed at a very high significance level (such as 0.03% for a firm that targets a AA rating) the VaR in (I.A.1.11) can be assessed at any percentile (in the example below we use 5%).

Returns and risks are usually estimated at a one-year time horizon on the assumption that the RAPM will help rank continuous activities rather than projects with a finite life. Often, financing costs are deducted from the numerator but the risk-free rate of return on the imputed economic

²⁵ See the Basel capital requirements, in particular the Basel II proposals (June 2004).

capital is added back. If the resulting RAPM exceeds the cost of equity capital of the firm, the activity is deemed to add value to the firm. In general, firms will seek to develop activities with high RAPM and curtail activities with low RAPM, especially those where the ratio is below the equity cost of the firm.

But basing business decisions on the RAPM rule may not produce the desired effect of increasing the value of the firm. Turnbull (2000) shows that RAROC measures, whilst taking into account a target probability of default, do not adjust for systematic risk and do not account properly for correlations between activities within a firm. He proposes instead an adjusted net present value rule that recognises marginal economic capital requirements and uses discount rates relevant for the risk of the underlying activity.

It remains the case that RAPMs may be useful in simpler circumstances such as the static portfolio selection problem when particular attention must be given to downside risks (but without setting a probability constraint on low returns).

I.A.1.7.4.2 Sortino Ratio, Omega Index and other Kappa indices

Back in 1959, Markowitz suggested the use of *semi-variance* rather than variance as a more relevant risk metric for investors worried about the downside. The semi-variance of a risky asset is defined as the expected value of the squared deviation of returns below the expected return. But Markowitz was discouraged by the computational difficulties involved in determining efficient portfolios in a return versus semi-variance framework. This is one of the difficulties with ‘downside’ risk metrics. The other is to ensure that the risk metric is based on a firm theoretical foundation (by establishing some compatibility with an acceptable utility function). As a general rule, downside risk metrics based on quantiles (such as VaR) or extremes are *not* compatible with utility theory and therefore are likely to lead to dubious results, whereas metrics based on lower partial moments *are* compatible with some utility functions.

The *lower partial moment* of order n ($n > 0$) of a random return R , given a threshold return τ is defined as:

$$\text{LPM}_n(\tau) = E[\max(\tau - R, 0)^n].$$

For example, the semi-variance is equal to $\text{LPM}_2(\mu)$ where $\mu = E[R]$. A lower partial moment can be used instead of a standard deviation in the denominator of the Sharpe ratio. This yields a family of *kappa indices* as follows:²⁶

$$K_n(\tau) = (\mu - \tau) / (\text{LPM}_n(\tau))^{1/n}. \quad (\text{I.A.1.12})$$

The second-order kappa index, introduced earlier by Sortino and van der Meer (1991), is called the *Sortino ratio*. The first-order kappa index is closely related to the *omega statistic* introduced by Keating and Shadwick (2002). Omega is the ratio of the expected return above threshold τ over the expected return below threshold τ , that is:

$$\Omega(\tau) = E[\max((R - \tau), 0)] / E[\max(\tau - R, 0)]. \quad (\text{I.A.1.13})$$

Kappa indices have the following properties:

- They decrease monotonically with the choice of threshold τ (they are all equal to zero when $\tau = \mu$).
- Their sensitivity to τ increases with the order, n .
- The kappa indices for portfolios with equal mean and variance are also sensitive to skewness (positive) and kurtosis (negative).
- The rankings of different assets under a kappa index depend on the choice of both n and τ ; large values of n and small values of τ emphasise the effects of skewness and kurtosis.

The definition of the kappa indices could also be extended to include a different return threshold for the calculation of excess return (numerator) and the calculation of lower partial moment (denominator). Currently there is much debate among fund managers and academics about the appropriate choice of parameters for kappa index with many recommending that these indices be calculated over a range of values for the parameters.²⁷

To illustrate the possible merits of downside RAPMs we return to the four gambles in Figure I.A.1.5. All four gambles have the same expected value of 5 and the same standard deviation of 15 and therefore the same Sharpe ratio of 1/3. Do downside RAPMs reveal preferences closer to what many observers feel intuitively, i.e. that A is the best investment followed by B, C and D in that order?

Table I.A.1.2 shows that the answer very much depends on which downside RAPM is used. But this result does not indicate which RAPM is most appropriate in which circumstance. For example, it is not clear which RAPM is the best metric for comparing the performance of hedge funds.

²⁶ So-called by their inventors, Paul Kaplan and J. Knowles (2003).

²⁷ We note that if the threshold for excess returns is taken as the risk-free rate, the kappa index will produce preference rankings congruent with the use of a *piecewise power-linear* utility function of the form:

$$\begin{aligned} u(x) &= x && \text{for } x \geq \tau, \\ u(x) &= x - c(x - \tau)^n && \text{for } x \leq \tau. \end{aligned}$$

Table I.A.1.2: Comparison of four investments according to four downside RAPMs

<p><i>ASR is defined in (I.A.1.9);</i> <i>Sortino is $K_2(0)$ as defined in (I.A.1.12);</i> <i>Omega is defined in (I.A.1.13) with a threshold of 0;</i></p>		ASR	Omega	Sortino	RoVaR
	Investment A	0.375	∞	∞	1.000
	Investment B	0.364	2.000	0.707	0.333
	Investment C	0.321	3.469	0.524	∞
	Investment D	0.276	2.250	0.395	0.111

I.A.1.8 Summary

We have surveyed the main risk preference criteria for making rational decisions under uncertainty. We now summarise the choice of criteria to be applied in specific circumstances and describe their likely effects compared to less formal approaches.

It may be helpful to visualise decisions under uncertainty in three dimensions:

- i. Size of risks: Are they large or small compared to net worth?
- ii. Portfolio effects: Are there dependencies among the risks we are taking, and in particular dependencies with existing risks (status quo) and external opportunities (securities markets), or can the new risks be considered independently of background risks?
- iii. Absolute levels or relative ranking: Is the problem to identify a class of efficient portfolios or do we have to decide on quantities, such as the size of an investment or level of insurance?

For the more complex cases involving large risks correlated with existing uncertainties (internal or external) and where decisions are about absolute levels of commitments (capital or other resources), there is no alternative to a full application of utility theory. A firm should have a corporate risk policy expressing risk preference in terms of a utility curve on its economic value (market value of equity or some more easily measurable proxy). This utility curve may be reduced to one or two parameters such as a risk-tolerance coefficient and the sensitivity of this coefficient to changes in value. The main effect of using a formal expression for risk attitude will be to improve the quality of communications and analyses concerning choices under uncertainty. It is likely that more projects deemed too risky at a departmental level will be recognised as

A sensible lower moment threshold τ is therefore the return level beyond which the investor is risk neutral. Below $x = \tau$ the coefficient of risk tolerance is proportional to $\tau - x$; a strange, endogenous utility function, but better than no utility representation at all.

valuable and passed on to higher decisional levels. It is also likely that the firm will benefit from greater consistency in risk taking across divisions and over time.

Where decisions are not so critical or complex, analytical shortcuts are possible. But one should remember that where decisions bear on absolute levels of risk, a value criterion must be used. The preference order it induces should still correspond to the preference order induced by a 'reasonable' utility function. For instance, the certain equivalent of a risky opportunity, measured by the expected value of its outcomes less a fraction of the incremental variance it creates, is a five-parameter measure that encapsulates the five essential ingredients: expected value and standard deviation of the new opportunity, standard deviation of existing risks, correlation between the new and the old risk, and trade-off between risk and return. It is the certain equivalent corresponding to an exponential utility function and jointly normally distributed risks; it is still a good approximation of a certain equivalent when the risk is not too large nor too skewed.

When the risks to be compared are not normally distributed, a value criterion based on symmetrical metrics such as the mean and variance may be misleading. Corrections can be made to include higher moments of the distributions or to focus on the more critical outcomes, for example the downside risks. But with powerful computers, one may doubt the wisdom of applying analytical approximations if it is at the expense of implying odd utility functions.

Where the problem at hand is not to decide on an allocation of resources but rather to compare performance, another simplification is possible: only the mildest assumptions about risk attitude, such as stochastic dominance, may be sufficient to compare performance. This is the classical portfolio selection problem where the issue is to identify an ideal portfolio or a family of so-called efficient portfolios. It is not a decision problem, only a first step towards making an investment decision. Under restrictive conditions, a ratio of excess return over the risk-free rate divided by the chosen risk metric can be justified as a relative performance criterion.

The Sharpe ratio is the grandfather of performance ratios and is still the most widely used. The risk metric in the denominator is the standard deviation of return. It is good for comparing the performance of assets with similarly distributed (two-parameter) returns, for instance some traded securities, but it does not apply to the comparison of activities that are not scalable, do not have similarly distributed returns, or have finite life cycles. The latter cases are not amenable to

the two-step approach: (i) identification of efficient portfolios through a relative performance measure and (ii) optimal investment in an efficient portfolio.²⁸

Alas, risk-adjusted performance measures— an elaboration on the concept of Sharpe ratio usually embedding a constraint on solvency risk and therefore relying on a lower quantile risk metric — have been applied to problems that only look superficially like portfolio selection of tradable securities. One should check whether applications of RAPMs are congruent with the widely claimed goal of maximising shareholder value; often they are not, especially if they do not account for systematic risks.

Finally, it may be worth noting that day-to-day risk management is mainly devoted to estimating expected losses rather than risks (see Chapter III.A.1). Many problems can be solved by an analysis of costs and benefits without resorting to a risk analysis. It should be possible to assume risk neutrality for all decisions where the range of consequences does not exceed a very small fraction of the net worth of a firm (say, 0.01% of capital for a bank, which would be €1 million for a €10 billion bank). The bar has been set lower than this in some financial institutions. The consequence is that too much management attention is given to minor risks, to the detriment of more significant risks. The management challenge is to identify the large risks and analyse them soundly.

²⁸ They may also necessitate a separate statement of time preference, as opposed to a simple discounting of cash flows at a certain cost of financing or target rate. The use of a higher discount rate to take into account the riskiness of future returns can be defended only in the case where returns are i.i.d. and a first-moment–second-moment criterion applies, because in those cases first and second moments are homogeneous in time. Otherwise it is misleading.

References

- Allais, M (1953) Le comportement de l'homme rationnel devant le risque, critique des postulats et axiomes de l'école américaine, *Econometrica*, 21, pp. 503–546.
- Artzner, P, Delbaen, F, Eber, J-M, and Heath, D (1999) Coherent measures of risk, *Mathematical Finance*, 9(3), pp. 203–228.
- Bernoulli, D (1738) Specimen theoriae novae de mensura sortis, *Commentarii Academiae Scientiarum Imperialis petropolitanae*, 5(2), pp. 175–192. Translated into English by L Sommer (1954) Exposition of a new theory on the measurement of risk, *Econometrica*, 22(1), pp. 23–26.
- Farquhar, P H (1984) Utility assessment methods, *Management Science*, 30, pp. 1283–1300.
- Fishburn, P (1970) *Utility Theory for Decision Making*. New York: Wiley.
- Kahneman, D, and Tversky, A (1979) Prospect theory: An analysis of decisions under risk, *Econometrica*, 47(2), pp. 263–291.
- Hodges, S D (1997) A generalisation of the Sharpe ratio and its applications to valuation bounds and risk measures. Working paper of the Financial Options Research Centre, University of Warwick.
- Jensen, M (1969) Risk, the pricing of capital assets, and the evaluation of investment portfolios, *Journal of Business*, 42(2), pp. 167–247.
- Keating, C, and Shadwick, F (2002) A universal performance measure, *Journal of Performance Measurement*, 6(3).
- Kaplan, P, and Knowles, J (2003) Kappa: A generalised downside-risk performance measure. Research paper, Morningstar:
http://datalab.morningstar.com/Midas/PDFs/KappaADownsideRisk_AdjustedPerformanceMeasure.pdf
- Kreps, D (1988) *Notes on the Theory of Choice*. Boulder CO: Westview.
- LiCalzi, M, and Sorato, A (2003) The Pearson system of utility functions. Working paper, Department of Applied Mathematics, University of Venice.
- Markowitz, H (1959) *Portfolio Selection*. New York: Wiley.
- Matten, C (2000) *Managing Bank Capital – Capital Allocation and Performance Measurement*, 2nd edition. New York: Wiley.
- Pratt, J W (1964) Risk aversion in the small and in the large, *Econometrica*, 32, pp. 122–136.

Ross, S (1976) The arbitrage theory of capital asset pricing, *Journal of Economic Theory*, 8, pp. 343-362.

Savage, L J (1954) *The Foundations of Statistics*. New York: Wiley. 2nd edition (1972), New York: Dover Publications, Inc.

Sharpe, W F (1964) Capital asset prices: A theory of market equilibrium under conditions of risk, *Journal of Finance*, 19, pp. 425–442.

Sortino, F, and van der Meer, R (1991) Downside risk, *Journal of Portfolio Management*, 17(4), pp. 27–31.

Spetzler, C S (1968) The development of corporate risk policy for capital investment decisions, *IEEE Transactions on Systems Science and Cybernetics*, **SSC-4**, pp. 279–300.

Treynor, J (1965) How to rate management of investment funds, *Harvard Business Review*, pp. 63–75.

Turnbull, S M (2000) Capital allocation and risk performance measurement in financial institutions, *Financial Markets, Institutions and Instruments*, 9(5).

von Neumann, J, and Morgenstern, O (1947) *The Theory of Games and Economic Behaviour*. Princeton, NJ: Princeton University Press.

Wakker, P, and Deneffe, D (1996) Eliciting von Neumann–Morgenstern utilities when probabilities are distorted or unknown, *Management Science*, 42, pp. 1131–1150.

Appendix I.A.1.A: Terminology

Everyday words carry many connotations, some of which may be unhelpful in the context of a specific theory. On the other hand, some nuances between similar words may be irrelevant for the purpose of a formal theory. But everyday words we must use lest we fall into technical jargon. When discussing decisions under uncertainty, no matter how serious these may be, we use the same formal language as when discussing games of chance. We do not mean to imply that life is a game in any moral sense or that there is no difference between investment and speculation in the business world. We are not addressing these issues and we do not want to disconcert or offend the reader with the use of apparently frivolous words. Hence we wish to clarify our use of the following terms.

Utility: The utility theory we discuss here has only historical connections with the now dated concept of economic utility and the rule of diminishing marginal utility, which applies to deterministic situations. It is also only remotely connected with the socio-philosophical school of utilitarianism. But utility is the word that has been used historically and was deemed good enough to be retained by von Neumann, Morgenstern, Savage and others. Therefore we keep it. We use many other words with a specific intent: uncertainty, risk, risk attitude, decisions, preferences, rationality, outcomes, gambles, lotteries, probabilities, etc. Many are defined in the text, some key terms are reviewed and compared here.

Outcome: A possible consequence of an action. An action may, of course, lead to an indefinite chain of outcomes and subsequent actions that would be infeasible to explore comprehensively. We focus our attention on outcomes that are material for the purpose of deciding between alternative courses of action.

Alternative words: event, consequence, result, reward, prize, profit, loss, penalty.

Gamble: A finite collection of exclusive and exhaustive outcomes that may result from a decision. Each outcome in a gamble has a certain probability of occurrence in the mind of a decision maker, and these probabilities sum up to one. The probabilities describe the uncertainty about which outcome will occur.

Alternative words: risky investment, lottery, wager, bet, risky opportunity, risky project, prospect.

Probability: A real number between zero and one expressing the degree of belief that a person attributes to the truth of a certain proposition; for example, the proposition that a particular outcome will result from playing a certain lottery. Probabilities in business situations are generally personal and conditional on a state of knowledge. Whether probabilities may be shared among individuals, or in some circumstances be objective, is not an issue: subjective and objective probabilities are manipulated in exactly the same way.

Alternative words: likelihood, chance.

Uncertainty: An incomplete state of knowledge about the truth of a proposition – in particular, about which among many possible outcomes may result from a certain decision. An uncertainty is described by a probability. We do not find it useful to distinguish risk (a more emotionally laden word, i.e., danger, and a narrower word, i.e., a concern for adverse deviations only) from uncertainty.²⁹

Alternative words: risk.

Alternative: A choice of two or more lotteries. Note that, in common parlance, the meaning of the word alternative (like those of the words choice and option) has slipped from our interpretation – a range from which to choose – to the thing being chosen, i.e., a lottery or a possibility.

Alternative words: game, decision problem, choice, option.

Decision: the irreversible, irrevocable selection of a particular lottery among a number of alternative lotteries. *Alternative words: act of choosing, course of action, selection, strategy.*

²⁹ In classical economics, some authors have made a distinction between risk and uncertainty according to whether objective probabilities could or could not be assigned to outcomes. This distinction is not relevant to decision making except that it may lead to an unfounded preference for objective probabilities, viz. the Ellsberg paradox, where a bet on the flip of a fair coin (50% 'objective' probability of winning) is preferred to a bet on the flip of a thumbtack (probability of winning that can only be better than or equal to 50% but is 'uncertain').

Appendix I.A.1.B: Utility Functions

Returning to the evaluation of large risks relative to the local coefficient of risk tolerance, we should make use of a full utility curve defined over the range of all possible outcomes. To obtain a complete curve we may proceed as in Section I.A.1.5, that is, generate a few points and then draw a freehand curve through these points. Alternatively, we may prefer to fit a known functional form to the empirical data. The main advantages of using a functional form are as follows:

- First, it ensures a smooth curve, that is, a curve without undesirable sharp kinks or changes in curvature as might result from a simple interpolation through a set of points. We have seen that the presence of kinks and curvature reversals is difficult to justify. It seems reasonable to aim for a curve exhibiting a smooth evolution of curvature, implying a smooth evolution of risk tolerance as wealth changes.
- Second, the choice of a functional form for a utility curve reduces the expression of risk attitude to the specification of a few parameters. The encoding of risk attitude is never very precise and it is more robust to specify a few parameters than to draw an entire curve empirically. It makes it easier to monitor and reassess the evolution of risk attitude over time. The functional form can also be selected to incorporate some desirable features such as constant absolute or constant relative risk aversion.
- Finally, a functional form is easier to communicate and to use than an empirical curve.

To illustrate these points, we consider three popular single-parameter utility functions (exponential, logarithmic and quadratic) and a two-parameter generalisation (the power utility function).

I.A.1.B.1 The Exponential Utility Function

This is given by:

$$u(x) = \lambda (1 - \exp(-x/\lambda)).$$

The argument x represents an increment to the current net worth of the firm; therefore the origin $x = 0$ should be interpreted as the current net worth.³⁰ We choose the utility scale so that $u(0) = 0$ and $u'(0) = 1$, hence, very small cash contributions ($x \ll \lambda$) make equal utility contributions. The curvature of the exponential utility function is constant, equal to $-1/\lambda$, hence our use of the coefficient of risk tolerance λ in the parameterisation. With a large λ , the utility function becomes almost linear and the firm exhibits near risk neutrality, whereas with a small λ , the firm exhibits a

large degree of risk aversion. If λ were to be negative, the firm would display a risk-seeking or gambling attitude, but that is almost impossible to imagine. A firm with a gambling behaviour, that is, a firm willing to pay more for risky projects than their expected value, would find it exceedingly difficult to raise any form of financing, except perhaps pure equity, and would be doomed to bankruptcy eventually.

The exponential utility curve has a key property that may appeal. If a sure quantity (positive or negative) is added to all outcomes of a risky project, the certain equivalent of the project, like its expected value, is increased by that quantity.³¹ Therefore the difference between the expected value and the certain equivalent – the risk discount – is independent of the current net worth. In particular, the certain equivalent, that is, the minimum selling price of a lottery, is the same as the maximum price a decision maker would be willing to pay to acquire it. We say that exponential utility functions display *constant absolute risk aversion* (CARA). Only exponential and linear utility functions (a degenerate form of exponential) share the CARA property. As a consequence of CARA, the current net worth of the firm is immaterial, and the origin of the outcome scale is arbitrary.

The effect of adding a constant to all outcomes of a gamble can also be extended to adding a second gamble, provided the second gamble is independent of the first; the certain equivalent of the two gambles is the sum of their respective certain equivalents. This is very convenient. It means that with an exponential utility function independent projects can be analysed separately. This property makes it easy to use an exponential form of utility function to delegate the decision making process. It also means that firms that are comfortable with adopting an exponential utility function will probably not have to review their coefficient of risk tolerance very often.

A final pleasant feature of exponential utility functions is that the calculation of an expected utility – and therefore of a certain equivalent, CE – can be obtained in closed analytical form for a number of well-known probability distributions. For example, the certain equivalent of a normally distributed portfolio return with mean $E[X]$ and variance $\text{Var}(X)$ under an exponential utility function with risk tolerance coefficient λ is simply:

$$\text{CE} = E[X] - \text{Var}(X)/2\lambda. \quad (\text{I.A.1.B.1})$$

³⁰ Net worth, or wealth, should be understood here in its broadest economic sense and not as some accounting definition of net asset value.

³¹ Proof: By definition of the certain equivalent of a gamble X , $u(\text{CE}(X)) = E[u(X)]$. With an exponential utility function, this leads to $\exp(-\text{CE}/\lambda) = \int \exp(-x/\lambda)p(x)dx$. If a fixed cash amount c is added to the gamble X then the right-hand side becomes $\int \exp(-(x+c)/\lambda)p(x)dx = \exp(-c/\lambda) \int \exp(-x/\lambda)p(x)dx = \exp(-(\text{CE}+c)/\lambda)$, that is, the certain equivalent of X is increased by c .

In other words, with exponential utility functions and normally distributed portfolio returns, the maximum EU principle reduces to the mean–variance criterion (I.A.1.2).

I.A.1.B.2 The Logarithmic Utility Function

This is given by:

$$u(x) = \lambda \ln(1 + x/\lambda).$$

Again, x stands for an incremental cash contribution to the current net worth and we choose the utility scale so that $u(0) = 0$ and $u'(0) = 1$.

The current local curvature (at $x = 0$) is also $u''/u' = -1/\lambda$, justifying the use of the local risk tolerance coefficient λ in the parameterisation. But the logarithmic form does not possess the CARA property of the exponential. The origin of the x scale has an absolute meaning; either we must move along the x -axis as new contributions are added to the firm's net worth or we must change the λ parameter. Indeed the λ parameter can be interpreted as a measure of the current net worth. Putting it another way, for $x = -\lambda$, the logarithmic utility becomes infinitely negative, and no losses beyond λ may ever be contemplated.³² Thus risk tolerance and net worth are tied together in this single parameter.

That is the feature that appealed to Bernoulli (1738). To quote him, 'any increase in wealth, no matter how insignificant, will always result in an increase in utility inversely proportionate to the quantity of goods already possessed.' In mathematical terms, if w represents total wealth, the first derivative $u'(w)$ is inversely proportional to w . Indeed, with the logarithmic utility function rewritten simply as $u(w) = \ln(w)$, we have $u'(w) = 1/w$.

Logarithmic utility functions share also a more general property known as *constant relative risk aversion* (CRRA) definable as follows: the local coefficient of risk tolerance varies linearly with wealth – in fact, it is equal to wealth. As we shall see in Section I.A.1.B.4, there is a more general class of utility functions, the power functions, for which the local coefficient of risk tolerance varies proportionally to wealth but the proportionality constant does not have to be equal to one, in fact it can be anything positive or negative.

³² Bernoulli (1738) defines net worth to include 'anything that can contribute to the adequate satisfaction of any sort of want. There is then nobody who can be said to possess nothing at all in this sense unless he starves to death'.

I.A.1.B.3 The Quadratic Utility Function

This is given by:

$$u(x) = x - (1/2\lambda)x^2.$$

Again, x stands for an incremental cash contribution to the current net worth and we choose the utility scale so that $u(0) = 0$ and $u'(0) = 1$ and the parameter λ is the current local coefficient of risk tolerance. The curvature is equal to $-1/(\lambda - x)$ and therefore the local risk aversion increases with x and becomes infinite when x reaches λ . This is odd. Even more strangely, utilities themselves start to decrease when x increases beyond λ , thus violating the principle of satiation. It may be advisable therefore to restrict the domain to $x < \lambda$ or to extend it with a constant $u(x) = \lambda/2$ for $x > \lambda$.

Yet, the quadratic utility function is sometimes used because of its simplicity: the expected utility of any gamble X under the quadratic utility function depends only on the expected value and variance of X . Therefore, it is easy to compute and the maximum EU principle reduces to an EV criterion. More generally, a quadratic utility function corresponds to the first three terms of a Taylor series expansion of any utility function. With the scaling and current local risk tolerance coefficient we chose, that is, with $u(0) = 0$, $u'(0) = 1$ and $u''(0) = -1/\lambda$, we indeed obtain:

$$u(x) = u(0) + u'(0)x + \frac{1}{2} u''(0)x^2 = x - (1/2\lambda)x^2.$$

I.A.1.B.4 The Power Utility Function

The power form of utility function, given by

$$u(x) = (\omega/(1 - \omega/\lambda))[(1 + x/\omega)^{1 - \omega/\lambda} - 1],$$

is scaled to produce $u(0) = 0$, $u'(0) = 1$ and $u''(0) = -1/\lambda$ as for the three previous utility functions. It therefore produces similar utilities for small risk. But it also contains a second parameter, ω , that can be used to control the variation of local risk tolerance with wealth.

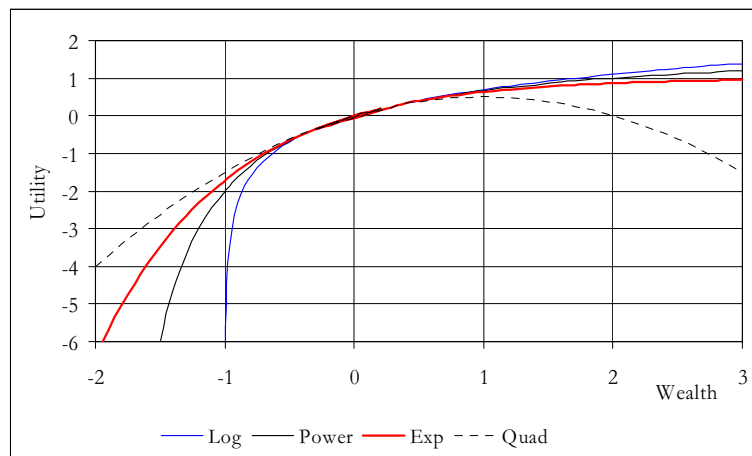
Its local risk tolerance is $-u'/u'' = \lambda(1 + x/\omega)$. If ω is positive, the local risk tolerance increases with wealth; conversely, if ω is negative, the local risk tolerance decreases when wealth increases. If ω approaches infinity, the risk tolerance tends towards a constant λ ; the power function converges towards an exponential utility function. If ω approaches λ it converges towards a logarithmic utility, and if $\omega = -\lambda$ we recover a quadratic utility.³³

Figures I.A.1.B.1 and I.A.1.B.2 display and contrast the behaviour of four members of the power family of utility functions and the corresponding local risk tolerances as a function of wealth. The figures show the three special cases when the power functions reduce to logarithmic, exponential and quadratic utilities. They also show an intermediate case between logarithmic and exponential utility functions with $\omega = 2\lambda$.

For many applications, it is sufficient to fit a power utility function to a few utility points obtained empirically. The two main parameters λ and ω have intuitive interpretations, λ being the current local coefficient of risk tolerance and λ/ω being the sensitivity of that coefficient to changes in wealth. Closed-form approximations for expected utilities and certain equivalents can be obtained as a function of λ , ω and the first few moments of the probability distribution of the outcome of a gamble. More general functional forms of utilities that can fit not only concave utilities but also the S-shaped forms of utilities suggested by prospect theory can be found in LiCalzi and Sorato (2003), together with a fitting procedure using Pearson's method of moments.

Figure I.A.1.B.1: Power utility family

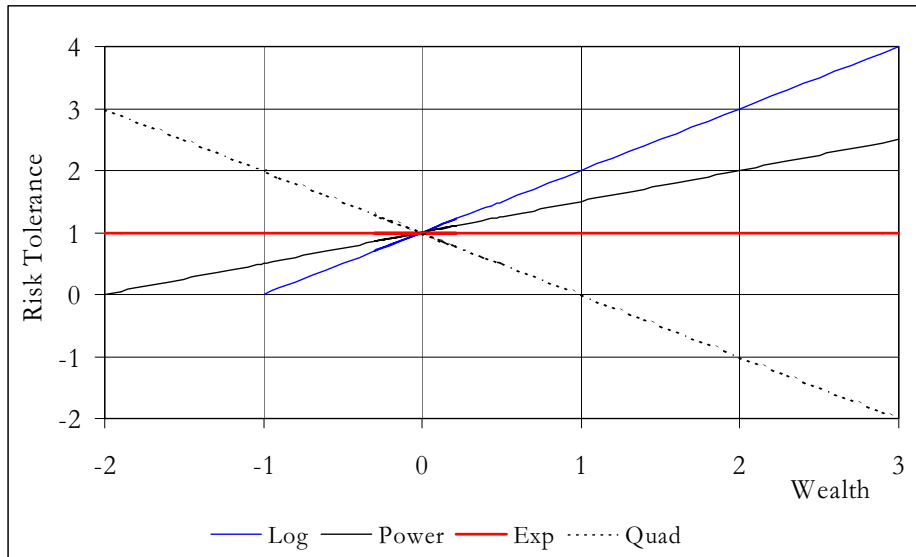
The four utility functions are members of the power utility family $u(x) = (\omega/(1 - \omega/\lambda))[(1 + x/\omega)^{1 - \omega/\lambda} - 1]$. They are all calibrated so that $u(0) = 0$, $u'(0) = 1$ and $u''(0) = -1/\lambda$. One obtains the logarithmic utility when ω tends towards λ ; the exponential when ω tends towards infinity and the quadratic when $\omega = -\lambda$. The scales for both wealth and utilities are in units of the risk tolerance coefficient λ . The fourth curve corresponds to $\omega = 2\lambda$.



³³ The class of power utility functions which contains the exponential, the logarithmic and the quadratic as special cases has been called the hyperbolic absolute risk aversion (HARA) class because the local absolute coefficient of risk aversion is hyperbolic in wealth: $\gamma = 1/\lambda(1+x/\omega)$.

Figure I.A.1.B.2: Risk tolerance with power utility family

The local risk tolerance is plotted for the four power utility functions in Figure I.A.1.B.1



I.A.2 Portfolio Mathematics

Paul Glasserman*

The main objective of this chapter is to explain how statistical parameters of portfolio returns — mean and variance — are determined by statistical parameters of the assets in the portfolio — their means, variances and correlations. In particular, this chapter emphasizes how correlations between the assets influence the variance of the portfolio's returns. The implications of this relationship are illustrated through various applications, including

- determining risk-return tradeoffs in alternative portfolio mixes;
- understanding how diversification reduces risk;
- finding the minimum-variance hedge ratio in hedging one asset with another;
- understanding how serial correlation in returns affects risk over different time horizons;
- calculating the value-at-risk in a portfolio;
- finding the probability that one portfolio outperforms another.

I.A.2.1 Means and Variances of Past Returns

Means and variances of asset returns can be understood in two different ways:

- as statistical descriptions of returns experienced in the past;
- as probabilistic predictions of returns to be experienced in the future.

In the first case, means and variances are calculated from historical data and serve to help summarize this data. In the second case, the mean (or expected value) and variance apply to random variables, not data. They help summarize the probability distribution of returns to be experienced in the future, rather than to summarize a record of past observations. Of course, we often use past data as a way to estimate parameters of the distribution of future returns, and in doing so we are connecting the two perspectives. It is nevertheless useful to keep the distinction in mind.

We begin by discussing means and variances of past returns because this case does not involve any probabilistic concepts or assumptions.

I.A.2.1.1 Returns

We use the notation x_1, x_2, \dots, x_n to denote n consecutive returns on some asset — e.g., monthly returns on a stock. Suppose, for example, that we have a record of stock prices S_0, S_1, \dots, S_n ; then the corresponding returns are the percentage price changes

$$x_i = \frac{S_i - S_{i-1}}{S_{i-1}}, \quad i = 1, \dots, n. \quad (\text{I.A.2.1})$$

*Columbia University, Graduate School of Business

Returns are sometimes computed on a continuously compounded basis, in which case

$$x_i = \ln(S_i/S_{i-1}). \quad (\text{I.A.2.2})$$

This simply states that

$$S_i = e^{x_i} S_{i-1}.$$

If the prices S_0, S_1, \dots, S_n are recorded at times t_0, t_1, \dots, t_n , then this means that the stock grew at a continuously compounded rate of $x_i/(t_{i+1} - t_i)$ over the period of time from t_{i-1} to t_i . With the first-order approximations

$$\ln(u) \approx u - 1, \quad e^x \approx 1 + x,$$

(I.A.2.1) and (I.A.2.2) become equivalent, so the two will be close, especially over short time horizons.

These equations assume that the stock pays no dividends. For a stock that pays dividends, the dividends should be included in the return calculation. Suppose the stock pays a dividend d_i at time t_i ; then (I.A.2.1) becomes

$$x_i = \frac{S_i + d_i - S_{i-1}}{S_{i-1}}, \quad i = 1, \dots, n.$$

If the dividend is paid sometime between t_{i-1} and t_i , then the d_i in this formula should include not only the dividend itself but also the income earned by reinvesting the dividend between the time it is paid and the date t_i .

Some assets may generate storage costs or other carrying costs; this is often true of commodities, for example. A carrying cost may be interpreted as a negative dividend for purposes of the return calculation.

I.A.2.1.2 Mean, Variance and Standard Deviation

From now on we assume that the returns x_1, \dots, x_n have been calculated over equally spaced intervals (e.g., daily returns or monthly returns) with proper accounting for dividends and any carrying costs.

The mean return, denoted by \bar{x} , is the arithmetic average

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i. \quad (\text{I.A.2.3})$$

The variance of the returns is the average squared difference from the mean,

$$s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (\text{I.A.2.4})$$

An alternative but algebraically equivalent formula is

$$s_x^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2.$$

Equation (I.A.2.4) defines what is usually called the *population* variance. The symbol s_x^2 is often reserved for the *sample* variance, given by

$$\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (\text{I.A.2.5})$$

The sample variance (I.A.2.5) is appropriate when we interpret x_1, \dots, x_n as a sample from a larger population of returns. The distinction between dividing by n and dividing by $n - 1$ is minor if n is even moderately large, so we will not dwell on this point.

In Microsoft Excel, the function VAR calculates (I.A.2.5). The function VARP calculates the population variance (I.A.2.4).

The standard deviation of the returns x_1, \dots, x_n is the square root of the variance,

$$s_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}. \tag{I.A.2.6}$$

This is the population standard deviation; the *sample* standard deviation is the square root of the sample variance (I.A.2.5). The Excel function STDEVP calculates (I.A.2.6), whereas STDEV calculates the sample standard deviation.

Variance and standard deviation are both measures of variability in a set of data: the more the individual values x_i deviate from the mean \bar{x} , the greater the values of s_x and s_x^2 .

As a consequence of the square root in (I.A.2.6), the standard deviation always has the same units as the underlying data x_1, \dots, x_n . If the data are percentages (as is often the case for returns), then the standard deviation is a percentage. If x_1, \dots, x_n were in dollars, then s_x would be in dollars, whereas the variance would be in dollars squared. This property often makes the standard deviation easier to interpret than the variance.

Example: Calculate the mean and standard deviation of the monthly returns listed in the first column of Table I.A.2.1.

Returns x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
-3.23%	-6.44%	0.00415
8.49%	5.28%	0.00279
4.20%	0.99%	0.00010
0.87%	-2.34%	0.00055
-12.41%	-15.62%	0.02440
8.56%	5.35%	0.00286
10.59%	7.38%	0.00545
4.35%	1.14%	0.00013
6.78%	3.57%	0.00127
4.83%	1.62%	0.00026
-2.33%	-5.54%	0.00307
7.82%	4.61%	0.00213
$\bar{x} = 3.21\%$		$s_x^2 = 0.00393$ $s_x = 6.27\%$

Table I.A.2.1: Calculation of mean, variance and standard deviation

The calculation is illustrated in the table. At the bottom of the first column, we have the mean 3.21%. The second column lists the deviations from the mean $x_i - \bar{x}$, and the third column lists the squared deviations $(x_i - \bar{x})^2$. The average of these squared deviations yields a variance of $s_x^2 = 0.00393$, and taking the square root yields a standard deviation of $s_x = 6.27\%$.

I.A.2.1.3 Portfolio Mean, Variance and Standard Deviation

Consider, now, a portfolio with holdings in two assets, XXX and YYY. Let $(x_1, y_1), \dots, (x_n, y_n)$ be pairs of past returns on the two assets. For example, if the returns are calculated on a monthly basis, then (x_1, y_1) are the returns on the two assets in the first month, (x_2, y_2) are the returns in the second month, and so on.

Consider a portfolio with weight w on XXX and weight $1 - w$ on YYY; e.g., $w = 0.6$ for a 60–40 weighting. What are the returns on the portfolio? Assuming the portfolio is rebalanced at the end of each period to maintain the same weights, the portfolio returns are given by the weighted asset returns,

$$\pi_i = wx_i + (1 - w)y_i, \quad i = 1, \dots, n. \quad (\text{I.A.2.7})$$

For example, if XXX increases by 5% and YYY increases by 10%, then the portfolio return is

$$(0.6)(5\%) + (0.4)(10\%) = 7\%.$$

How are the mean, variance and standard deviation of the portfolio returns related to those of the assets in the portfolio? The mean portfolio return is just the weighted average of the individual asset returns:

$$\bar{\pi} = w\bar{x} + (1 - w)\bar{y}.$$

This can be verified by noting that the mean portfolio return

$$\bar{\pi} = \frac{1}{n} \sum_{i=1}^n \pi_i = \frac{1}{n} \sum_{i=1}^n (wx_i + (1 - w)y_i)$$

is equal to the weighted average

$$w \left(\frac{1}{n} \sum_{i=1}^n x_i \right) + (1 - w) \left(\frac{1}{n} \sum_{i=1}^n y_i \right) = w\bar{x} + (1 - w)\bar{y}.$$

This is illustrated in Table I.A.2.2 for a portfolio holding 60% XXX and 40% YYY. The monthly portfolio returns in the last column are calculated month by month from returns on the two assets. The average of the portfolio returns in the last column is 2.73%. This is the same as the value obtained by taking the weighted average

$$0.6\bar{x} + 0.4\bar{y} = (0.6)(3.21\%) + (0.4)(2.02\%) = 2.73\%$$

of the mean returns on the two assets.

In contrast, the variance and standard deviation of the portfolio are not quite as simply related to those of the assets in the portfolio. This is illustrated by the example in Table I.A.2.2, which shows the variances and standard deviations calculated from each column of numbers using the formulas (I.A.2.4) and (I.A.2.6). The portfolio standard deviation is not given by the weighted average of the asset standard deviations,

$$s_{\pi} \neq ws_x + (1 - w)s_y,$$

nor is the portfolio variance given by the weighted average of the asset variances,

$$s_{\pi}^2 \neq ws_x^2 + (1 - w)s_y^2.$$

To relate the portfolio variance and standard deviation to the individual assets, we need to introduce the correlation between the asset returns.

	XXX Returns	YYY Returns	Portfolio Returns
	–3.23%	3.29%	–0.62%
	8.49%	3.99%	6.69%
	4.20%	8.84%	6.06%
	0.87%	–3.27%	–0.79%
	–12.41%	–7.98%	–10.64%
	8.56%	6.51%	7.74%
	10.59%	2.82%	7.48%
	4.35%	3.38%	3.96%
	6.78%	5.13%	6.12%
	4.83%	2.48%	3.89%
	–2.33%	–2.20%	–2.28%
	7.82%	1.20%	5.17%
Mean	3.21%	2.02%	2.73%
Variance	0.00393	0.00193	0.00268
Std Dev	6.27%	4.39%	5.18%

Table I.A.2.2: Calculation of portfolio returns and their mean, variance and standard deviation

I.A.2.1.4 Correlation

Correlation measures the direction and strength of the relationship between two sets of observations x_1, \dots, x_n and y_1, \dots, y_n . The sign of the correlation determines the direction, with a positive correlation meaning that above average values of one variable tend to be paired with above average values of the other, and negative correlation meaning that above average values of one variable tend to be paired with below average values of the other. The magnitude of the correlation measures the strength of the association — more precisely, it measures the strength of the *linear* relationship between the two variables, with perfect correlation meaning that the points (x_i, y_i) lie exactly on a straight line.

To define correlation precisely, we first introduce the covariance between the x_i and the y_i , given by

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}). \tag{I.A.2.8}$$

The Excel function COVAR applied to two ranges containing the x_i and y_i executes this calculation.

Observe that s_{xy} will be large and positive if those i for which $x_i > \bar{x}$ coincide with those for which $y_i > \bar{y}$. Conversely, s_{xy} will be large and negative when values $x_i > \bar{x}$ are paired with values $y_i < \bar{y}$.

A shortcoming of covariance as a measure of association is that it has no natural scale: what counts as a “large” covariance depends on context and even on the units of the data. If the x_i are measurements in meters, for example, then changing their units to centimeters would increase the covariance by a factor of 100 without in any way changing the relationship between the x_i and y_i .

The correlation coefficient gets around this shortcoming by using the standard deviations of the observations to normalize the covariance:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}. \tag{I.A.2.9}$$

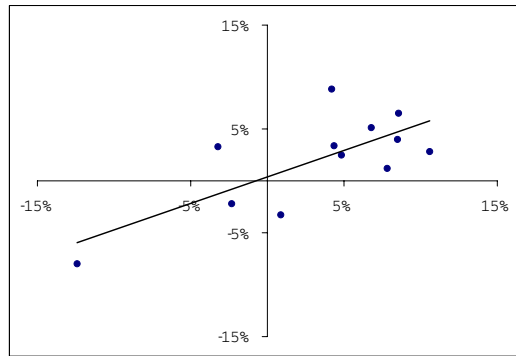


Figure I.A.2.1: Scatter plot of the monthly returns (x_i, y_i) from Table I.A.2.2 and the regression line through the data.

Here, s_x is the standard deviation in (I.A.2.6) and s_y is the standard deviation of y_1, \dots, y_n . The Excel function CORREL applied to two ranges containing the x_i and y_i calculates (I.A.2.9).

The correlation coefficient r_{xy} has the same sign as the covariance s_{xy} , so the interpretation of the sign as an indicator of positive or negative association between two variables is the same as before. But the correlation coefficient satisfies

$$-1 \leq r_{xy} \leq 1,$$

and is thus on a universal scale, regardless of the underlying observations (x_i, y_i) or their units.

The extreme cases $r_{xy} = \pm 1$ occur if and only if the points fall on a straight line, meaning that

$$y_i = a + bx_i, \quad i = 1, \dots, n,$$

for some intercept a and slope b . If $b > 0$, then $r_{xy} = 1$ and if $b < 0$ then $r_{xy} = -1$.

At intermediate values of r_{xy} , the magnitude $|r_{xy}|$ gives an indication of how well the points (x_i, y_i) are approximated by a straight line. If the points (x_i, y_i) form a cloud, with no evident pattern, the correlation coefficient will be close to zero.

However, a correlation of zero does not necessarily indicate the absence of a relation between the two variables — it indicates the absence of a *linear* relation. For example, if we take for the x_i the integers $-5, -4, \dots, 4, 5$ and set $y_i = x_i^2$, then $r_{xy} = 0$ though there is evidently a very strong relation between the x_i and y_i . This strong relation just happens not to be a linear relation.

Figure I.A.2.1 shows a scatter plot of the monthly returns (x_i, y_i) in Table I.A.2.2: each pair (x_i, y_i) provides the coordinates of one point in the figure. The figure also shows the regression line through the data, which is the best linear fit in the least-squares sense. Using (I.A.2.9), we can calculate a correlation of $r_{xy} = 0.7287$ for the data in Table I.A.2.2. Consistent with this value, the figure indicates a positive association between the x_i and y_i , and only a moderately strong linear relation between the two.

The magnitude of the correlation coefficient measures how closely the regression line fits the data, but it should not be confused with the slope of the regression line, which is given by

$$b = \frac{s_y}{s_x} r_{xy}.$$

The line in Figure I.A.2.1 has a slope of 0.51.

I.A.2.1.5 Correlation and Portfolio Variance

We now return to the portfolio returns in Table I.A.2.2 and the question we posed earlier about the connection between the portfolio variance s_π^2 and the variances of the individual assets. Recall that the portfolio returns π_i are related to the asset returns x_i, y_i through the weights w and $1 - w$ as in (I.A.2.7). Having introduced correlation, we can now present one of the most important identities of this chapter:

$$s_\pi^2 = w^2 s_x^2 + (1 - w)^2 s_y^2 + 2w(1 - w)s_x s_y r_{xy}. \quad (\text{I.A.2.10})$$

Thus, the variance of the portfolio returns depends not only on the variances s_x^2 and s_y^2 of the individual assets but also on the correlation between them.

Using (I.A.2.9), we can alternatively write (I.A.2.10) as

$$s_\pi^2 = w^2 s_x^2 + (1 - w)^2 s_y^2 + 2w(1 - w)s_{xy} \quad (\text{I.A.2.11})$$

which uses the covariance s_{xy} in place of $s_x s_y r_{xy}$.

Equation (I.A.2.10) is an algebraic identity that results from the definitions in (I.A.2.4), (I.A.2.7) and (I.A.2.9). It is useful because it decomposes the portfolio variance into simpler quantities — asset variances and correlation. In so doing, it reveals which features of the asset returns are relevant to the variance of the portfolio returns.

In (I.A.2.10) we see that a positive correlation $r_{xy} > 0$ will lead to a larger value of the portfolio return s_π^2 , and a negative correlation will lead to a smaller value.

This fits with intuition. If the two asset returns are negatively correlated, then losses in one asset will tend to be offset by gains in the other, resulting in lower variability in the overall portfolio returns.

But it must be stressed that diversification — splitting investments across more than one asset — reduces variance even if the assets are positively correlated. We return to this point in Section I.A.2.2.

Example: Apply (I.A.2.10) to the returns in Table I.A.2.2. The correlation between the monthly returns x_i and y_i in the table is 0.7287. This can be calculated using (I.A.2.9) or, more easily, by applying the Excel function CORREL to the data. Using the values $s_x = 6.27\%$ and $s_y = 4.39\%$ calculated previously (see the bottom of the table) and recalling the weights $w = 0.6$, $1 - w = 0.4$, the right side of (I.A.2.10) becomes

$$(0.6)^2(0.0627)^2 + (0.4)^2(0.0439)^2 + 2(0.6)(0.4)(0.0627)(0.0439)(0.7287) = 0.00268$$

and thus coincides with the value computed in the table by taking the variance of the individual returns π_1, \dots, π_n .

I.A.2.1.6 Portfolio Standard Deviation

By taking the square root of each side of (I.A.2.10), we find that the portfolio standard deviation s_π is given by

$$s_\pi = \sqrt{w^2 s_x^2 + (1 - w)^2 s_y^2 + 2w(1 - w)s_x s_y r_{xy}}. \quad (\text{I.A.2.12})$$

For any weight w , $0 < w < 1$, the portfolio standard deviation satisfies

$$s_\pi \leq w s_x + (1 - w) s_y,$$

with strict inequality unless $r_{xy} = 1$. Thus, diversification reduces the portfolio standard deviation (compared to the weighted average of the asset standard deviations) except in the extreme case $r_{xy} = 1$. When $r_{xy} = 1$, the returns of the two assets are in a straight line relation with each other, so investing in one is equivalent to investing in the other.

In Section I.A.2.6, we will see that correlation is also relevant to the problem of annualizing standard deviations calculated from daily or monthly returns. The relevant correlation parameter in that setting is the correlation between returns of a single asset (or portfolio) in different time periods, rather than the correlation between returns of different assets in the same time period. We will see, for example, that in the absence of serial correlation an annual standard deviation is $\sqrt{12}$ times larger than a monthly standard deviation — not 12 times larger. This may be interpreted as the result of diversification over time.

I.A.2.2 Mean and Variance of Future Returns

The discussion in the previous section applies to data (x_i, y_i) , $i = 1, \dots, n$, such as observations of past returns. In particular, the relationships (I.A.2.10) and (I.A.2.12) relating the portfolio variance and standard deviation to properties of the assets in the portfolio are purely algebraic and do not rely on any probabilistic or statistical assumptions. As such, these relationships are merely descriptive of historical data.

To turn from a historical perspective to a forward-looking one, we need to introduce a probabilistic formulation. We will model future returns as random variables and show that analogous relationships between the portfolio and individual assets hold in this case as well.

Most of the formulas in this section are counterparts of those in the previous section. The key difference is that in the previous section we gave equal weight to all returns, but now in calculating a forward-looking mean or variance we will weight each possible return by its probability.

I.A.2.2.1 Single Asset

Denote the (unknown and uncertain) return on an asset over the next month by the random variable X . The random variable is characterized by its distribution. If X can take only a finite number of possible values — as in a model that considers a finite number of possible scenarios — its distribution can be specified through a probability mass function p_X , for which

$$p_X(x) = P(X = x).$$

Here, $P(\cdot)$ denotes the probability of the event in parentheses, so $p_X(x)$ is the probability that the random variable X takes the value x . Alternatively, the distribution of X may be specified through a probability density f_X (as in the case of normally distributed returns), in which case the probability that the random variable X is less than or equal to an arbitrary value x is given by

$$P(X \leq x) = \int_{-\infty}^x f_X(u) du.$$

We denote the expected value of the random variable by $E[X]$ or μ_X and also refer to it as the mean of the random variable. For a random variable taking only the values x_1, \dots, x_n , the expected value is the probability-weighted average of these values and is given by

$$E[X] = \mu_X = \sum_{i=1}^n x_i p_X(x_i).$$

For a random variable with a probability density, the expected value is given by

$$E[X] = \mu_X = \int_{-\infty}^{\infty} x f_X(x) dx.$$

We denote the variance of the random variable by $\text{Var}[X]$ or σ_X^2 . It is given by

$$\text{Var}[X] = \sigma_X^2 = \sum_{i=1}^n (x_i - \mu_X)^2 p_X(x_i)$$

in the case of a mass function, and by

$$\text{Var}[X] = \sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx$$

in the case of a probability density. In both cases, the standard deviation, $\text{StdDev}[X]$ or σ_X , is given by the square root of the variance.

The expressions in (I.A.2.3) and (I.A.2.4) coincide with what one obtains for the mean and variance of a random variable that is equally likely to take any of the values x_1, \dots, x_n .

I.A.2.2.2 Covariance and Correlation

Now let the random variables X and Y denote the unknown returns of two assets over the next period. Their joint distribution may be specified through a mass function

$$p_{XY}(x, y) = P(X = x \text{ and } Y = y)$$

giving the joint probability that X takes the value x and Y takes the value y . The joint distribution may alternatively be given by a joint density f_{XY} . In this case, the probability that $X \leq x$ and $Y \leq y$ is

$$P(X \leq x \text{ and } Y \leq y) = \int_{-\infty}^x \int_{-\infty}^y f_{XY}(u, v) du dv.$$

The random variables are independent if $p_{XY}(x, y) = p_X(x)p_Y(y)$ for all x, y , in the first case, or if $f_{XY}(x, y) = f_X(x)f_Y(y)$ in the second case. Intuitively, independence means that the value of one random variable has no information about the value of the other.

The covariance between X and Y is given by

$$\text{Cov}[X, Y] = E[(X - \mu_X)(Y - \mu_Y)], \tag{I.A.2.13}$$

the expectation calculated using the joint distribution of X and Y .

The correlation between X and Y , denoted by ρ_{XY} , is given by

$$\rho_{XY} = \frac{\text{Cov}[X, Y]}{\sigma_X \sigma_Y}.$$

The correlation coefficient ρ_{XY} for random variables should be interpreted in much the same way as the coefficient r_{xy} for data points: it measures the strength of the linear relationship between X and Y .

I.A.2.2.3 Mean and Variance of a Linear Combination

Let the random variables X and Y denote the unknown returns of two assets over the next period. The return on a portfolio with weight w on the first asset and weight $1 - w$ on the second is given by the random variable

$$\Pi = wX + (1 - w)Y.$$

In order to present expressions for the mean and variance of the portfolio return Π , we first consider the more general case of an arbitrary linear combination $aX + bY$, with constants a and b not necessarily summing to 1. For the mean of the linear combination, we have simply

$$E[aX + bY] = aE[X] + bE[Y] = a\mu_X + b\mu_Y, \quad (\text{I.A.2.14})$$

a linear combination of the means.

For the variance, the counterpart to (I.A.2.10) is

$$\text{Var}[aX + bY] = a^2\text{Var}[X] + b^2\text{Var}[Y] + 2ab\text{Cov}[X, Y].$$

Equivalently,

$$\text{Var}[aX + bY] = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_X\sigma_Y\rho_{XY}. \quad (\text{I.A.2.15})$$

The standard deviation is obtained by taking the square root.

In the special case $a = w$ and $b = 1 - w$, we get the portfolio mean

$$\mu_\Pi = w\mu_X + (1 - w)\mu_Y, \quad (\text{I.A.2.16})$$

the portfolio variance

$$\sigma_\Pi^2 = w^2\sigma_X^2 + (1 - w)^2\sigma_Y^2 + 2w(1 - w)\sigma_X\sigma_Y\rho_{XY}, \quad (\text{I.A.2.17})$$

and the portfolio standard deviation is given by the square root of (I.A.2.17).

I.A.2.2.4 Example: Portfolio Return

Suppose the annual returns on XXX and YYY have the following characteristics:

	Mean	Std Dev	
XXX	$\mu_X = 12\%$	$\sigma_X = 24\%$	$\rho_{XY} = 0.33$
YYY	$\mu_Y = 15\%$	$\sigma_Y = 30\%$	

(I.A.2.18)

Find the expected return and the standard deviation of the return of a portfolio invested 60% in XXX and 40% in YYY.

The portfolio return is $\Pi = 0.6X + 0.4Y$, which has the form in (I.A.2.14) and (I.A.2.15). Thus, the expected return is

$$E[\Pi] = 0.6\mu_X + 0.4\mu_Y = (0.6)(12\%) + (0.4)(15\%) = 13.2\%. \quad (\text{I.A.2.19})$$

The expected portfolio return of 13.2% lies between the expected returns (12% and 15%) of the individual assets.

The variance of the portfolio return is

$$\text{Var}[\Pi] = 0.6^2\sigma_X^2 + 0.4^2\sigma_Y^2 + 2(0.6)(0.4)\sigma_X\sigma_Y\rho_{XY} = 0.04654, \quad (\text{I.A.2.20})$$

using (I.A.2.15) and the values of σ_X , σ_Y , and ρ_{XY} displayed above. By taking the square root of the variance, we get the portfolio standard deviation

$$\sigma_\Pi = \text{StdDev}[\Pi] = \sqrt{0.04654} = 21.57\%. \quad (\text{I.A.2.21})$$

Notice that this is smaller than the standard deviations $\sigma_X = 24\%$ and $\sigma_Y = 30\%$ of the individual assets. Thus, diversification has achieved a reduction in variability, *even though the asset returns are positively correlated*.

I.A.2.2.5 Example: Portfolio Profit

We continue to use the asset return characteristics of the previous example. Consider a \$10,000 investment in XXX. What is the expected profit from this investment over one year? What is the standard deviation of the profit?

If the return on XXX over one year is given by the random variable X , then each dollar invested in XXX grows to $1 + X$ dollars over the course of the year. (Here we are assuming that the return X is a simple return, in the sense of (I.A.2.1), rather than a continuously compounded return.) It follows that the profit on a \$10,000 investment is given by

$$\Pi = aX, \quad a = 10,000.$$

This has the form $aX + bY$, with $b = 0$. The expected profit is

$$\text{E}[aX] = a\mu_X = (10,000)(12\%) = 1,200.$$

The standard deviation is

$$\text{StdDev}[aX] = \sqrt{\text{Var}[aX]} = \sqrt{a^2\sigma_X^2} = (10,000)(24\%) = 2,400.$$

Because the mean and standard deviation always have the same units as the random variable, the units are now dollars rather than percentage points.

Now consider a short position of \$5,000 invested in YYY. Find the expected profit and the standard deviation of the profit.

The profit is

$$\Pi = bY, \quad b = -5,000.$$

The expected profit is

$$\text{E}[bY] = b\mu_Y = (-5000)(15\%) = -750.$$

The standard deviation is

$$\text{StdDev}[bY] = \sqrt{\text{Var}[bY]} = \sqrt{b^2\sigma_Y^2} = (5000)(30\%) = 1500.$$

Notice, more generally, that

$$\text{StdDev}[bY] = |b|\sigma_Y. \quad (\text{I.A.2.22})$$

because the standard deviation is the square root of $b^2\sigma_Y^2$, and the square root of b^2 is the positive number $|b|$, regardless of the sign of b . This confirms the intuitively evident fact that the standard deviation of a short position is the same as the standard deviation of the opposite long position. The fluctuations in the two positions are mirror images of each other; this affects the sign of the expected profit, but not the level of variability.

I.A.2.2.6 Example: Long and Short Positions

We now combine the long and short positions of the previous example. Consider an investment of \$10,000 in XXX and a short position of \$5,000 in YYY. Find the mean and standard deviation of the portfolio profit.

The profit is

$$\Pi = aX + bY = 10,000X - 5,000Y.$$

Its expected value is

$$E[aX + bY] = a\mu_X + b\mu_Y = (10,000)(12\%) + (-5000)(15\%) = 1,200 - 750 = 450.$$

To get the standard deviation, we first compute the variance, which is

$$\begin{aligned} a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_X\sigma_Y\rho_{XY} = \\ 10000^2(0.12^2) + 5000^2(0.15^2) - 2(10000)(5000)(0.12)(0.15)(0.33) = 5,634,000. \end{aligned} \quad (\text{I.A.2.23})$$

By taking the square root, we get a standard deviation of \$2,374.

Notice that in (I.A.2.23) the correlation term has a negative coefficient because a and b have opposite signs. This reflects the fact that if the two assets are positively correlated, then taking opposite positions in the two assets has the effect of offsetting fluctuations in one asset with fluctuations in the other.

I.A.2.2.7 Example: Correlation

Find the correlation between the long and short positions in the previous example.

This is a special case of the more general problem of finding the correlation between aX and bY . By definition (see (I.A.2.9)) the correlation is

$$\frac{\text{Cov}[aX, bY]}{\text{StdDev}[aX]\text{StdDev}[bY]}.$$

From (I.A.2.13) we find that

$$\text{Cov}[aX, bY] = ab\text{Cov}[X, Y].$$

From (I.A.2.22) we know that

$$\text{StdDev}[aX] = |a|\sigma_X, \quad \text{StdDev}[bY] = |b|\sigma_Y.$$

Thus, the correlation between aX and bY is

$$\frac{ab\text{Cov}[X, Y]}{|a|\sigma_X|b|\sigma_Y} = \frac{ab}{|a||b|}\rho_{XY}.$$

This shows that the correlation between aX and bY has the same magnitude as the correlation between X and Y , though it may have different sign. More specifically, if a and b have the same sign, then the correlation between aX and bY is ρ_{XY} ; if a and b have opposite signs, then the correlation between aX and bY is $-\rho_{XY}$. Thus, the two parts of the portfolio in the previous example have a correlation of -0.33 , whereas XXX and YYY are positively correlated. The correlation between the two positions in the portfolio becomes negative because the portfolio is long XXX and short YYY.

This conclusion is consistent with our earlier interpretation of correlation as a measure of the strength of the linear relation between two variables: aX and bY are neither more nor less linearly related than X and Y , but the direction of the relation depends on the signs of a and b .

I.A.2.3 Mean–Variance Tradeoffs

In (I.A.2.19)–(I.A.2.20), we calculated the mean and variance of a portfolio return with a fixed set of portfolio weights. We now investigate how these features of the portfolio return vary as we vary the portfolio weights w and $1 - w$. We continue to use the same values for means, variances and correlation of the asset returns as in (I.A.2.18). The weight w is the fraction of the portfolio invested in XXX.

We begin by posing two questions:

- What are the achievable portfolio return means as w varies between 0 and 1?
- What are the achievable portfolio return variances as w varies between 0 and 1?

I.A.2.3.1 Achievable Expected Returns

At $w = 0$, the portfolio is fully invested in YYY and its expected return is $\mu_Y = 15\%$. At $w = 1$, the portfolio is fully invested in XXX and its expected return is $\mu_X = 12\%$. Between these extremes, the portfolio mean,

$$\mu_P = w\mu_X + (1 - w)\mu_Y$$

varies linearly with w . So, for any value of w between 0 and 1, the portfolio’s expected return will lie proportionately between the two extremes. For example, in (I.A.2.19) we found that at $w = 60\%$, the expected portfolio return is 13.2%.

Similarly, for any value of μ_Π between 12% and 15% we can find a value of w that achieves μ_Π as its expected return. We do this by solving for w in the equation

$$\mu_\Pi = w\mu_X + (1 - w)\mu_Y$$

to get

$$w = \frac{\mu_Y - \mu_\Pi}{\mu_Y - \mu_X} = \frac{15\% - \mu_\Pi}{15\% - 12\%}.$$

Thus, by varying w between 0 and 1 we can attain any value of μ_Π between 12% and 15%, and these are the only attainable values of μ_Π with w in this range. This is illustrated in Figure I.A.2.2.

I.A.2.3.2 Achievable Variance and Standard Deviation

The behavior of the variance is less straightforward. The portfolio variance is given by

$$\sigma_\Pi^2 = w^2\sigma_X^2 + (1 - w)^2\sigma_Y^2 + 2w(1 - w)\sigma_X\sigma_Y\rho_{XY}.$$

At the extremes of $w = 0$ and $w = 1$, we get the individual asset variances σ_Y^2 and σ_X^2 . Between these extremes, the portfolio variance is a *quadratic* function of w .

The portfolio standard deviation,

$$\sigma_\Pi = \sqrt{w^2\sigma_X^2 + (1 - w)^2\sigma_Y^2 + 2w(1 - w)\sigma_X\sigma_Y\rho_{XY}}$$

varies similarly. At $w = 0$, we have $\sigma_\Pi = \sigma_Y = 30\%$ and at the other extreme, $w = 1$, we have $\sigma_\Pi = \sigma_X = 24\%$. Every value between 24% and 30% is achievable by some value of w

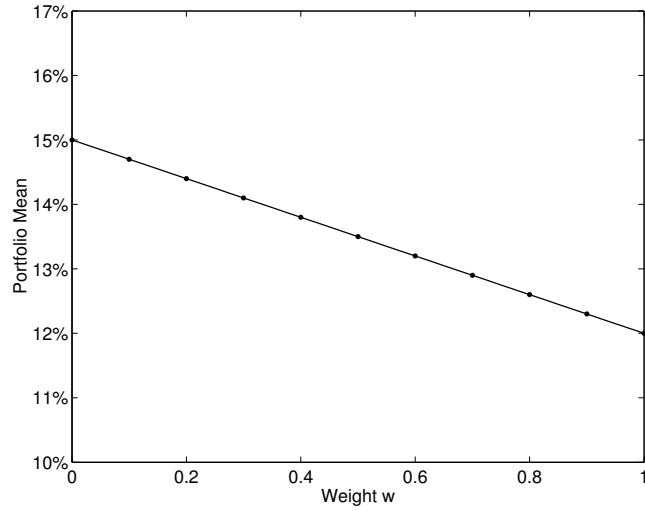


Figure I.A.2.2: Portfolio mean as function of the fraction w invested in XXX

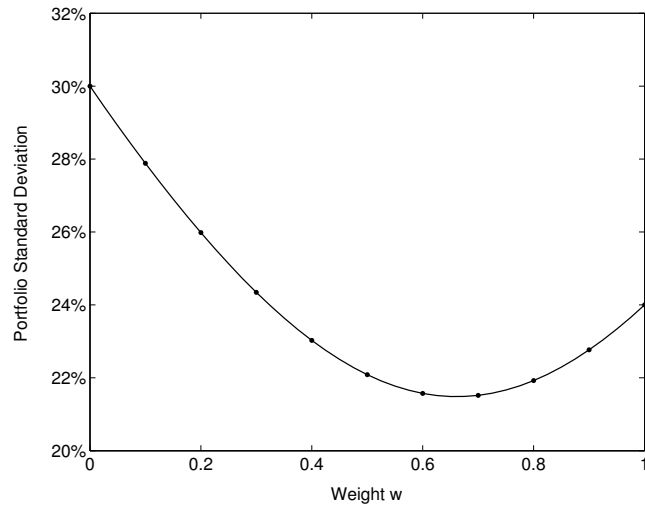


Figure I.A.2.3: Portfolio standard deviation as function of the fraction w invested in XXX

between 0 and 1, *but these are not the only achievable values*. We can also attain values of σ_{Π} strictly smaller than 24%. This is illustrated in Figure I.A.2.3, which shows that the portfolio standard deviation dips below 24% as w increases from 0 to 1.

The minimum variance and the minimum standard deviation occur at the same value of w , because if one portfolio has a smaller variance than another, then it also has a smaller standard deviation. We can find the minimizing weight w by differentiating σ_{Π}^2 , setting the derivative equal to zero, and solving for w ; i.e., solving the equation

$$w\sigma_X^2 - (1-w)\sigma_Y^2 + \sigma_X\sigma_Y\rho_{XY} - 2w\sigma_X\sigma_Y\rho_{XY} = 0.$$

As indicated by Figure I.A.2.3, the minimum occurs not at one of the extremes, but rather at

$$w = \frac{\sigma_Y^2 - \sigma_X\sigma_Y\rho_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_X\sigma_Y\rho_{XY}} = 0.662. \quad (\text{I.A.2.24})$$

Thus, even though XXX has lower standard deviation than YYY, the portfolio with the smallest standard deviation is not the XXX-only portfolio with $w = 1$. We can achieve a strictly smaller standard deviation of $\sigma_{\Pi} \approx 21.5\%$ by investing 66.2% in XXX and 33.8% in YYY.

I.A.2.3.3 Achievable Combinations of Mean and Standard Deviation

Figure I.A.2.4 combines the information in Figures I.A.2.2 and I.A.2.3 to show the combinations of portfolio mean and standard deviation that can be attained by varying the weight w .

Each point on the curve corresponds to a value of w and thus to a particular portfolio. The YYY-only portfolio ($w = 0$) has a mean of $\mu_Y = 15\%$ and a standard deviation of $\sigma_Y = 30\%$; it corresponds to the upper-right-most point in the figure. The XXX-only portfolio ($w = 1$) has a mean of $\mu_X = 12\%$ and a standard deviation of $\sigma_X = 24\%$; it corresponds to the lowest point in the figure. As we vary w between 0 and 1, we get the corresponding combinations of mean and standard deviation by moving along the curve. For example, at $w = 0.2$, we get a portfolio standard deviation of 26% and a mean of 14.4%.

The portfolio with the smallest standard deviation corresponds to the leftmost point on the curve in Figure I.A.2.4. As calculated in (I.A.2.24), the minimum standard deviation is achieved at $w = 0.662$. The resulting expected return is 13%.

I.A.2.3.4 Efficient Frontier

Different investors may have different preferences and may therefore choose different combinations of risk and return, which here we interpret as different combinations of portfolio mean and standard deviation. A very risk-averse investor would choose a portfolio with weight w near 0.662, since this is the value that minimizes the standard deviation. An investor with a greater appetite for risk would prefer a portfolio closer to the upper-right corner of Figure I.A.2.4 (near $w = 0$) as these are the points with highest expected return and highest standard deviation.

We cannot say that a portfolio with $w = 0.662$ is better than one with $w = 0$, or vice-versa. Different investors may reasonably prefer either of these portfolios to the other.

We can, however, say that a rational investor should consider only those portfolios in the upper portion of the curve, indicated by the thicker portion of the line. This is one of the

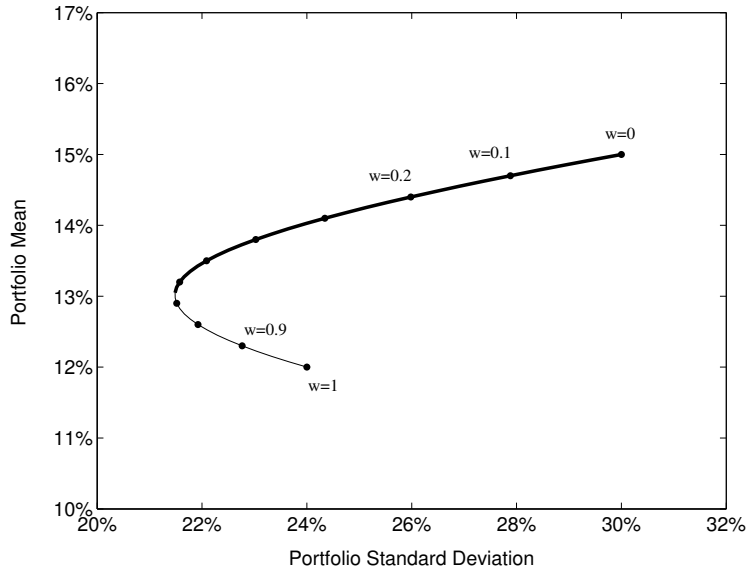


Figure I.A.2.4: The curve shows the combinations of portfolio mean and standard deviation that can be attained by varying the fraction w invested in XXX. The thick portion of the curve shows the efficient frontier.

key insights of Markowitz’s theory of portfolio selection.¹ Rationality here means preferring a higher expected return and/or a lower standard deviation, other things being equal.

According to this criterion, no rational investor should consider a portfolio with w greater than 0.662. Consider, for example, the portfolio with $w = 0.9$. It has an expected return of about 12.3% and a standard deviation of about 22.8%. At the same level of risk, the investor could get a higher expected return by moving vertically in Figure I.A.2.4 to a higher point on the curve. The vertical line defined by a standard deviation of 22.8% crosses the portfolio curve twice; see Figure I.A.2.5. At the higher intersection, the investor gets an expected return of about 13.7%, without taking on any additional risk. This intersection (labeled with a circle in Figure I.A.2.5) occurs at $w \approx 0.425$. All investors should prefer this portfolio to the one with $w = 0.9$.

Indeed, from Figure I.A.2.5 we see that every portfolio with $0.9 < w \leq 0.425$ should be preferred to the portfolio with $w = 0.9$ by all investors. All portfolios in this range have higher mean and lower standard deviation.

We say that a portfolio is *dominated* if there is at least one other portfolio with higher mean and lower standard deviation. In a graph that plots mean against standard deviation, a portfolio is dominated if another portfolio lies to its northwest. The portfolio with $w = 0.9$ is thus an example of a dominated portfolio. The undominated portfolios in Figures I.A.2.4 and I.A.2.5 are the ones indicated by the thick portion of the curve. This portion of the set of achievable portfolios is called the *efficient frontier*.

I.A.2.3.5 Utility Maximization

We have noted that a rational investor should consider only those portfolios on the efficient frontier, but that different investors may prefer different combinations of risk and return along

¹Markowitz, H. (1952) Portfolio selection. *Journal of Finance*, 7:77–91.

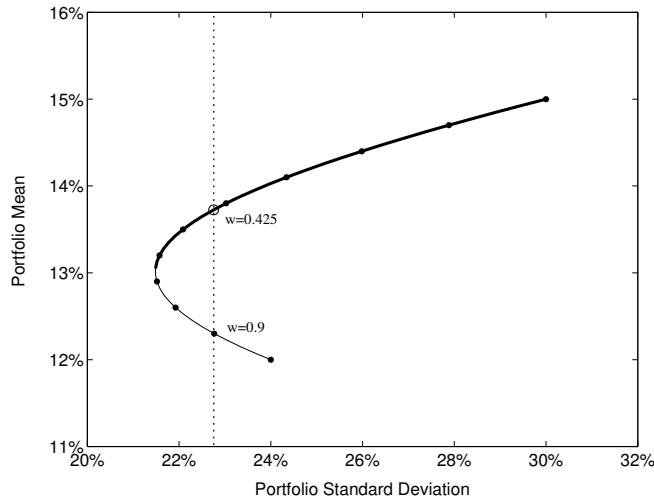


Figure I.A.2.5: The portfolio with $w = 0.425$ has higher expected return and the same standard deviation as the portfolio with $w = 0.9$. The portfolio with $w = 0.9$ is therefore dominated.

the efficient frontier.

A simple model of the tradeoff between risk and return posits that each investor selects a portfolio by maximizing a utility function of the form

$$\mu_{\Pi} - \frac{\gamma}{2}\sigma_{\Pi}^2. \tag{I.A.2.25}$$

Here, μ_{Π} and σ_{Π}^2 are the portfolio mean and variance, and $\gamma \geq 0$ is a measure of risk aversion: the greater the value of γ , the more the utility function penalizes risk. The investor maximizes utility by choosing a value of the portfolio weight w , $0 \leq w \leq 1$; this then determines μ_{Π} and σ_{Π}^2 .

At $\gamma = 0$, the investor is indifferent to risk and seeks to maximize expected return. The corresponding optimum is $w = 0$, producing the YYY-only portfolio. As γ increases without bound, the investor becomes increasingly risk-averse, in the limit choosing the minimum-variance portfolio, $w = 0.66$. As γ varies between 0 and ∞ , the resulting optimal weights w sweep out the efficient frontier.

I.A.2.3.6 Varying the Correlation Parameter

Figures I.A.2.4 and I.A.2.5 are based on the parameters for assets XXX and YYY given in (I.A.2.18). In particular, the correlation is $\rho_{XY} = 0.33$ for the returns on the two assets.

Figure I.A.2.6 illustrates the effect of varying the correlation coefficient while holding the other parameters ($\mu_X, \mu_Y, \sigma_X, \sigma_Y$) fixed. Each curve in the figure shows the achievable pairs of means and standard deviations under a different value of ρ_{XY} . The curve labeled $\rho = 0.33$ coincides with the original curve in Figure I.A.2.4.

Varying ρ has no effect on the set of achievable means: as explained previously, the possible portfolio means are all the values between the two asset means of 12% and 15%.

Varying ρ does, however, have a pronounced effect on the achievable standard deviations and the benefit of diversification. As we move from $\rho = 0.33$ to $\rho = 0$ and then $\rho = -0.5$, we

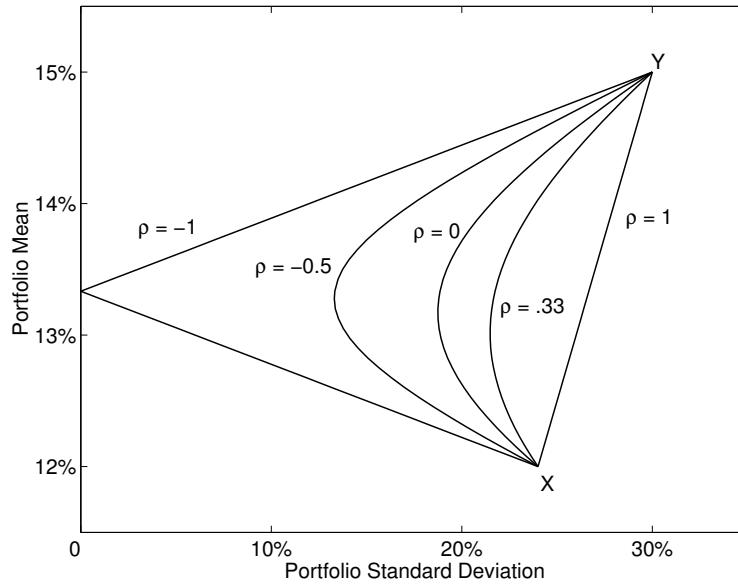


Figure I.A.2.6: Achievable pairs of mean and standard deviation for varying levels of the correlation between the two assets in the portfolio.

see that diversifying the portfolio has a greater impact on risk reduction. In particular, the smallest possible standard deviation (the leftmost point on each curve) decreases as the assets becomes less positively correlated and then more negatively correlated.

At the extreme of $\rho = -1$, perfect negative correlation, variability in one asset can be used to completely offset variability in the other asset and thus to create a portfolio with a standard deviation of zero — a riskless portfolio.

The only case in which we find no diversification benefit is $\rho = 1$, perfect positive correlation. At this extreme, the curve of achievable portfolios is a straight line between the XXX-only and YYY-only portfolios. Because the returns of the two assets move in lock-step at this extreme, it is not possible to offset variability in one with a long position in the other. As a consequence, the smallest achievable standard deviation is 24%, the standard deviation for XXX.

I.A.2.4 Multiple Assets

We now explain how the ideas of the previous section extend to portfolios of more than two assets.

To consider multiple assets, we need to introduce some notation. We let d denote the number of assets and let X_1, X_2, \dots, X_d denote their (unknown, random) returns over the next period. The returns have means

$$\mu_i = E[X_i], \quad i = 1, \dots, d,$$

and variances

$$\sigma_i^2 = \text{Var}[X_i], \quad i = 1, \dots, d.$$

The covariance between X_i and X_j is denoted by

$$\sigma_{ij} = \text{Cov}[X_i, X_j].$$

The covariance can also be expressed as

$$\sigma_{ij} = \sigma_i \sigma_j \rho_{ij},$$

with ρ_{ij} the correlation between X_i and X_j , and σ_i, σ_j their respective standard deviations.

I.A.2.4.1 Portfolio Mean and Variance

Now consider a portfolio invested in these d assets with weight w_i on the i th asset, $i = 1, \dots, d$. This means that a fraction w_i , $0 \leq w_i \leq 1$, of the portfolio's value is held in asset i . We assume that $w_1 + w_2 + \dots + w_d = 1$.

The resulting portfolio return is described by the random variable

$$\Pi = w_1 X_1 + w_2 X_2 + \dots + w_d X_d.$$

The expected portfolio return is

$$\mu_\Pi = \mathbf{E}[\Pi] = w_1 \mu_1 + w_2 \mu_2 + \dots + w_d \mu_d,$$

and the variance of the portfolio return is

$$\sigma_\Pi^2 = \text{Var}[\Pi] = \sum_{i=1}^d w_i^2 \sigma_i^2 + 2 \sum_{i=1}^d \sum_{j=i+1}^d w_i w_j \sigma_i \sigma_j \rho_{ij}. \quad (\text{I.A.2.26})$$

The portfolio standard deviation, σ_Π , is the square root of the variance.

I.A.2.4.2 Vector-Matrix Notation

Formulas for portfolios of multiple assets can be written compactly using vector and matrix notation. By default, we take all vectors to be column vectors and use the symbol \top to indicate the transpose.

We denote the asset returns, expected returns and portfolio weights by the vectors

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix}.$$

The portfolio return is given by the scalar product

$$\Pi = \mathbf{w}^\top \mathbf{X}.$$

The portfolio's expected return is

$$\mu_\pi = \mathbf{w}^\top \boldsymbol{\mu}.$$

Let $\boldsymbol{\Sigma}$ denote the variance-covariance matrix of the returns X_1, \dots, X_d . This is the $d \times d$ matrix with entries

$$\boldsymbol{\Sigma}_{ij} = \sigma_i \sigma_j \rho_{ij}, \quad i, j = 1, \dots, d.$$

Along the diagonal $i = j$, we have $\rho_{ii} = 1$ (each asset is perfectly correlated with itself), so this reduces to

$$\boldsymbol{\Sigma}_{ii} = \sigma_i^2,$$

the variance of the i th asset return.

The portfolio variance can be expressed compactly as

$$\sigma_\Pi^2 = \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}$$

using $\boldsymbol{\Sigma}$ and the portfolio weights \mathbf{w} .

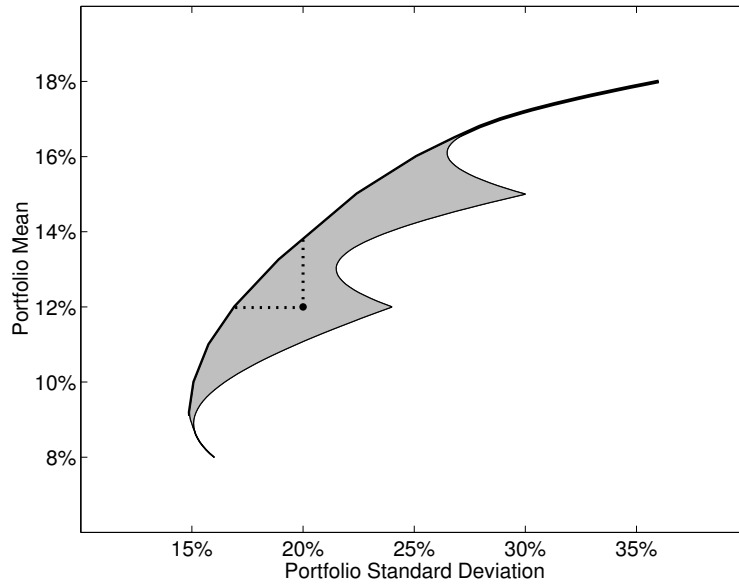


Figure I.A.2.7: The shaded region shows achievable pairs of mean and standard deviation for a portfolio invested in four assets. The thick line marking the upper boundary of the region is the efficient frontier.

I.A.2.4.3 Efficient Frontier

By varying the portfolio weights w_1, \dots, w_d , we get different combinations of the portfolio mean μ_Π and standard deviation σ_Π . As in the case of two assets, different investors may choose different combinations, but there are some combinations that no rational investor should choose because they are dominated by other portfolios.

To illustrate this point, we consider a four-asset example. The individual asset means and standard deviations are given by

$$(\mu_1, \mu_2, \mu_3, \mu_4) = (8\%, 12\%, 15\%, 18\%), \quad (\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (16\%, 24\%, 30\%, 36\%).$$

For simplicity, we take all the correlations $\rho_{ij}, j \neq i$, equal to 0.33. These parameters determine the set of pairs (σ_Π, μ_Π) that can be achieved by varying portfolio weights w_1, w_2, w_3 and w_4 .

Achievable pairs of portfolio mean and standard deviation for this example are shown in Figure I.A.2.7. The four kink points on the right side of the region correspond to the four underlying assets: the risk–return combinations that can be attained by fully investing the portfolio in a single asset are $(16\%, 8\%), (24\%, 12\%), (30\%, 15\%)$ and $(36\%, 18\%)$.

The curves connecting the kink points are the combinations that can be achieved by combining just those two assets. For example, the curve that connects the two middle kink points reproduces the curve in Figure I.A.2.4, because the second and third assets in this example have the same parameters as the assets XXX and YYY used in Figure I.A.2.4.

By investing the portfolio in all four assets, we can achieve all the combinations in the shaded region in Figure I.A.2.7. This shows that the possibilities are much richer with four assets than they were with just two.²

²The shaded region shows only those combinations that are at least as attractive as combinations that can be achieved with just two of the four assets. One can construct inferior combinations — lying to the right of the shaded region — using the four assets, but these have been omitted from the figure.

The thick line marking the upper boundary of the shaded region indicates the efficient frontier. These are the only combinations a rational investor should consider. Any combination not on the efficient frontier is *dominated*, meaning that there are other portfolios that achieve higher expected returns for the same risk or lower risk for the same expected returns.

The point marked by a dot inside the shaded region of Figure I.A.2.7 is an example of a dominated portfolio. Indeed, all combinations to the northwest (i.e., in the region bounded by the dotted lines) dominate because they offer a higher mean and a lower standard deviation.

The portfolios on the efficient frontier correspond to solutions to the problem

$$\begin{aligned} & \min_{\mathbf{w}} \mathbf{w}^\top \Sigma \mathbf{w} \\ & \text{subject to } \mathbf{w}^\top \boldsymbol{\mu} = \mu_* \\ & 0 \leq w_i \leq 1, i = 1, \dots, d \\ & w_1 + w_2 + \dots + w_d = 1, \end{aligned}$$

as μ_* varies between the lowest and highest means of the underlying assets. This is a quadratic programming problem. It asks for the minimum variance that can be achieved with an expected return of μ_* .

The efficient frontier can also be characterized as the set of solutions to the utility maximization problem with utility function (I.A.2.25), as the risk aversion parameter γ varies.

I.A.2.5 A Hedging Example

As a further application of the ideas developed in the previous sections, we now consider a problem of hedging commodity price risk.

I.A.2.5.1 Problem Formulation

FlyFreight, a cargo company, has won a contract for a major shipping job that will begin in three months. As part of this job, FlyFreight will need to buy two million gallons of jet fuel.

In order to win the contract, FlyFreight bid aggressively, and it is concerned that an increase in jet fuel prices could wipe out its profit on the deal. It therefore decides to try to hedge its price risk with futures contracts. As there are no exchange-traded contracts on jet fuel, it follows the industry practice of hedging with heating oil contracts traded on the New York Mercantile Exchange. Heating oil and jet fuel are chemically quite similar and their price fluctuations have thus historically shown a high degree of correlation.

The spot prices of heating oil and jet fuel are currently at 65 cents and 68 cents per gallon, respectively. The heating oil futures price for delivery in three months is 67 cents per gallon. Each heating oil futures contract is for 42,000 gallons (1,000 barrels).

Key to the hedging strategy is the observation that if the price of heating oil rises, then so does the value of a heating oil futures contract. When FlyFreight enters in a futures contract at 67 cents per gallon, the contract has zero value. If the price of heating oil rises to 70 cents, the contract — which allows the holder to buy at 67 cents — acquires positive value. Moreover, FlyFreight can collect this positive value by closing out its position, without ever buying or selling heating oil. If the price of heating oil drops to 65 cents, then the contract — which still commits the holder to buying at 67 cents — acquires negative value, and FlyFreight must pay to close out its position. Thus, the value of the contract moves in the same direction as the futures price of heating oil, and this will usually be the same direction as the price of jet fuel.

We would like to address the following questions:

- (i) What is the price risk (as measured by standard deviation) faced by FlyFreight if it does no hedging?
- (ii) What is the price risk if FlyFreight hedges jet fuel with heating oil on a gallon-for-gallon basis?
- (iii) What is the risk-minimizing hedge? How effective is the best hedge?

To address these questions, we need more information. Suppose the percentage change (the “return”) in the price of jet fuel over three months has a standard deviation of 18%. The standard deviation of the percentage change in the futures price of heating oil over three months is 23%. The correlation between the two is 0.82.

In order to quantify the risk faced by FlyFreight, we will work with dollar amounts rather than percentage changes. So, let

$$\begin{aligned} X &= \text{change in heating oil futures price per gallon over three months, and} \\ Y &= \text{change in jet fuel price per gallon over three months.} \end{aligned}$$

Then, based on the information above, we find that X has a standard deviation of

$$\sigma_X = 23\% \times \$0.67 = \$0.1541$$

and Y has a standard deviation of

$$\sigma_Y = 18\% \times \$0.68 = \$0.1224.$$

We further posit that X and Y have zero expected value. This is particularly appropriate in the case of X , because a futures price already reflects expectations about price changes over the life of the contract. In contrast, the expected value for a change in the spot price of jet fuel could reasonably be different from zero. In this case, we could reinterpret Y as the difference between the spot price in three months and the expected price (which would imply $E[Y] = 0$). For example, if the expected change in the price of jet fuel is 2 cents per gallon, we could assume that this expected price increase was reflected in FlyFreight’s bid; the risk to which FlyFreight is exposed lies, then, in deviations from this expected increase, not in the increase itself.

With this formulation of the problem, we can answer question (i), above. The change in FlyFreight’s unhedged position over the next three months is

$$-2,000,000Y,$$

because FlyFreight will need to buy 2,000,000 gallons and each gallon will cost Y more than expected. This position has a standard deviation (see (I.A.2.22)) of

$$2,000,000\sigma_Y = \$244,800.$$

I.A.2.5.2 Gallon-for-Gallon Hedge

Consider, next, a simple hedging strategy in which FlyFreight hedges the 2 million gallons of jet fuel it needs with 2 million gallons of heating oil. As each heating oil contract is for 42,000 gallons, this means that FlyFreight goes long

$$\frac{2,000,000}{42,000} = 47.6 \text{ contracts,}$$

or simply 48 contracts.

FlyFreight will not take delivery of heating oil. Instead, it will close out its positions before the futures expire. If the prices of both heating oil and jet fuel go up, FlyFreight’s costs will go up but this will be offset, at least partly, by a profit from the heating oil contracts. If both prices go down, FlyFreight will lose money on the futures, but its cargo contract will be more profitable because its fuel costs will be lower.

With a gallon-for-gallon hedge, FlyFreight’s position thus becomes

$$(48)(42,000)X - 2,000,000Y.$$

This is actually a slight simplification in that it ignores the difference between futures and forwards.

Using (I.A.2.15), we find that the standard deviation of this position is

$$\sqrt{(48 \times 42,000)^2 \sigma_X^2 + (2,000,000)^2 \sigma_Y^2 - 2(48)(42,000)(2,000,000) \sigma_X \sigma_Y \rho_{XY}}.$$

With $\sigma_X = 0.1541$, $\sigma_Y = 0.1224$, and $\rho_{XY} = 0.82$, we get

$$\text{StdDev}[(48)(42,000)X - 2,000,000Y] = \$178,092,$$

which is lower than the unhedged standard deviation of \$244,800.

I.A.2.5.3 Minimum-Variance Hedge

Can FlyFreight do better? Instead of assuming a gallon-for-gallon hedge, we now treat the number of contracts c as a variable over which we will optimize.

With c contracts, FlyFreight’s position becomes

$$42,000cX - 2,000,000Y.$$

The variance of the position is therefore

$$(42,000c)^2 \sigma_X^2 + (2,000,000)^2 \sigma_Y^2 - 2(42,000c)(2,000,000) \sigma_X \sigma_Y \rho_{XY} \quad (\text{I.A.2.27})$$

and the standard deviation is the square root of this expression.

The variance of the position is a quadratic function of c . We can minimize the variance by differentiating with respect to c , setting the derivative equal to zero, and solving for c . In other words, we need to solve

$$(42,000)^2 \sigma_X^2 c - (42,000)(2,000,000) \sigma_X \sigma_Y \rho_{XY} = 0.$$

The solution is

$$c^* = \left(\frac{2,000,000}{42,000} \right) \frac{\sigma_Y}{\sigma_X} \rho_{XY}. \quad (\text{I.A.2.28})$$

This evaluates to

$$c^* = \left(\frac{2,000,000}{42,000} \right) \left(\frac{0.1224}{0.1541} \right) 0.82 = 31.0$$

This is the number of contracts that minimizes the variance (equivalently, the standard deviation) of the hedged position.

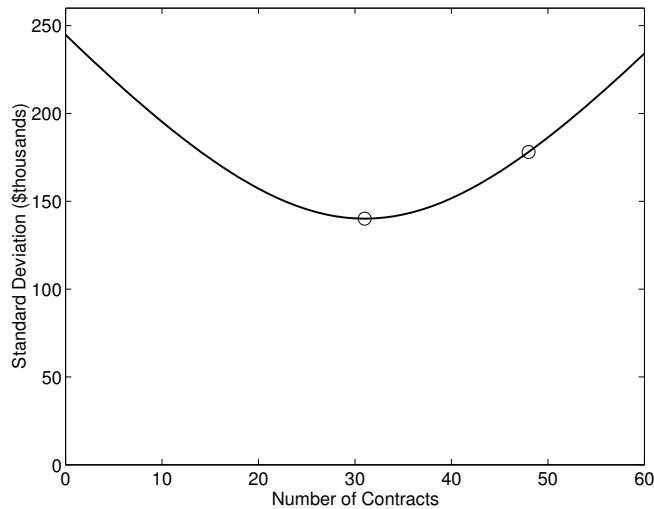


Figure I.A.2.8: Standard deviation as a function of the number of contracts used to hedge.

Figure I.A.2.8 plots the standard deviation against the number of contracts. The figure confirms that hedging with 48 contracts reduces risk compared to no hedge and that the optimal hedge is 31 contracts.

The shape of the curve in Figure I.A.2.8 follows from the fact that the variance (I.A.2.27) is quadratic in the number of contracts, but it can also be explained more intuitively as follows. If we start at $c = 0$ and increase the number of contracts, the risk initially decreases because fluctuations in jet fuel prices are partly offset by fluctuations in heating oil prices; this is the effect of the last term in (I.A.2.27), which is negative. However, beyond a certain point, the additional risk resulting from a long position in heating oil begins to overwhelm the hedging effect — the first term in (I.A.2.27) becomes greater than the third. In this “overhedging” region, the variance and standard deviation increase as we increase the number of contracts. The optimal point $c^* = 31$ is the point that balances these two effects.

It is useful to decompose the optimal number of contracts in (I.A.2.28) into two pieces. The most important part is

$$\beta = \frac{\sigma_Y}{\sigma_X} \rho_{XY} \tag{I.A.2.29}$$

which evaluates to

$$\beta = \left(\frac{0.1224}{0.1541} \right) 0.82 = 0.65.$$

This is the minimum-variance hedge ratio for hedging a gallon of jet fuel with gallons of heating oil. If we had exactly one gallon of jet fuel to hedge, we would hedge it with β gallons of heating oil. For 2,000,000 gallons of jet fuel we need $2,000,000\beta$ gallons of heating oil. To convert this from gallons to contracts, we divide by 42,000, the size of each contract. That gives c^* in (I.A.2.28).

Notice that β does not equal 1, the coefficient implicit in the gallon-for-gallon hedging strategy. We would have $\beta = 1$ if the two price changes had the same standard deviation and were perfectly correlated.

I.A.2.5.4 Effectiveness of the Optimal Hedge

We have seen that the risk-minimizing hedge for FlyFreight requires 31 contracts. But how effective is this optimal hedge? By how much does it reduce FlyFreight’s risk?

One way to answer this question is to look at the ratio

$$\frac{\text{StdDev}[\text{Optimal hedge}]}{\text{StdDev}[\text{No hedge}]}$$

of the standard deviation with the optimal hedge and with no hedge. We have already found that the denominator is \$244,080. To find the numerator, we can substitute $c = 31$ in (I.A.2.27) and then take the square root of the resulting value, which gives \$140,115. The reduction in standard deviation is thus

$$\frac{140,115}{244,080} = 57.24\%. \tag{I.A.2.30}$$

Equivalently, we may say that the optimal hedge eliminates 42.76% of the standard deviation.

It is instructive to look at this calculation in greater generality. In calculating the optimal number of contracts c^* , we minimized an expression of the form

$$\sigma^2(c) = a^2c^2\sigma_X^2 + b^2\sigma_Y^2 - 2abc\sigma_X\sigma_Y\rho_{XY},$$

with $a, b > 0$. The optimal value of c is

$$c^* = \frac{b\sigma_Y}{a\sigma_X}\rho_{XY}.$$

It follows by substituting the values $c = c^*$ and $c = 0$ that

$$\frac{\sigma(c^*)}{\sigma(0)} = \sqrt{1 - \rho_{XY}^2}.$$

In other words,

$$\boxed{\frac{\text{StdDev}[\text{Optimal hedge}]}{\text{StdDev}[\text{No hedge}]} = \sqrt{1 - \rho_{XY}^2}} \tag{I.A.2.31}$$

At $\rho_{XY} = 0.82$, we get

$$\sqrt{1 - \rho_{XY}^2} = 0.5724,$$

consistent with what we obtained previously in (I.A.2.30).

The interesting feature of the risk reduction in (I.A.2.31) is that it is completely determined by the correlation coefficient ρ_{XY} . This means that from the outset, FlyFreight could have looked at the correlation $\rho_{XY} = 0.82$ between heating oil and jet fuel and known immediately that the optimal hedge would reduce risk to 57.24% of the unhedged risk.

Equation (I.A.2.31) also indicates that the effectiveness of the optimal hedge depends on the magnitude of the correlation, but not its sign. If the two commodities had a correlation of -0.82 , we would achieve exactly the same hedge effectiveness, though of course in this case the optimal hedge would sell rather than buy heating oil contracts.

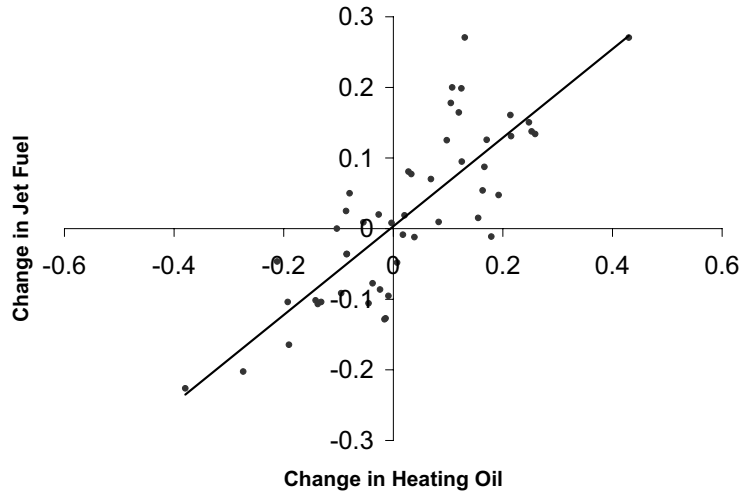


Figure I.A.2.9: Scatter plot of price changes, with regression line

I.A.2.5.5 Connection with Regression

The minimum-variance hedge ratio β in (I.A.2.29) is the slope of the regression line in regressing Y against X . This leads to a graphical interpretation of minimum-variance hedging.

Figure I.A.2.9 shows a (hypothetical) scatter plot of changes in jet fuel prices and changes in the futures price of heating oil over three-month intervals. The standard deviations and correlation of the points in the figure match the values given above for σ_X , σ_Y and ρ_{XY} . The line through the figure is the least-squares regression line.

The slope of the line, β , measures the average change in the price per gallon of jet fuel for each unit change in the future price per gallon of heating oil. The slope is thus an indication of how one asset moves with the other.

Consider, for example, an extreme case in which the changes fall exactly on a straight line:

$$\Delta\text{Jet Fuel Price} = \beta\Delta\text{Heating Oil Futures Price};$$

i.e.,

$$Y = \beta X.$$

In this case, a position of the form $Y - \beta X$ would be riskless, as would a position of the form

$$-2,000,000Y + 2,000,000\beta X. \tag{I.A.2.32}$$

This is, in fact, the minimum-variance hedge in (I.A.2.28), because

$$2,000,000\beta = 2,000,000 \times 0.65 = 1,300,000 \text{ gallons}$$

which is 31 contracts.

In Figure I.A.2.9, the linear relation does not hold exactly. Instead we have

$$Y = \beta X + \epsilon$$

for some residual ϵ . The optimal hedge (I.A.2.32) now leaves an unhedged variability represented by

$$-2,000,000Y + 2,000,000\beta X = 2,000,000\epsilon.$$

The residual in the regression is the portion of the variability in Y that cannot be removed by hedging with X . The optimal risk reduction (I.A.2.31) is

$$\frac{\text{StdDev}[\epsilon]}{\text{StdDev}[Y]} = \sqrt{1 - \rho_{XY}^2}.$$

The connection between minimum-variance hedging and regression is particularly useful in hedging with multiple assets. Let Y denote the price change to be hedged and let X_1, \dots, X_d denote the changes in prices of the hedging instruments. We can use regression to estimate coefficients in the equation

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_d X_d + \epsilon.$$

The variance-minimizing hedge is then

$$Y - \beta_1 X_1 - \dots - \beta_d X_d.$$

In other words, to hedge a long position in Y , we should sell β_i units of the i th asset for each unit of Y (or buy $|\beta_i|$ units if $\beta_i < 0$). The effectiveness of this hedge (in the sense of the risk-reduction ratio (I.A.2.31)) is given by $\sqrt{1 - R^2}$, where R^2 is the usual coefficient of determination in the regression.

I.A.2.6 Serial Correlation

In Section I.A.2.4, we explained how correlations between pairs of assets affect the variability of portfolio returns. We now show how similar ideas can be used to relate the variability in returns in a single asset over different time horizons. For example, we can address the following question:

- How is the standard deviation of annual returns of a single asset related to the standard deviation of monthly returns for the same asset?

Throughout this section, we use X_1, \dots, X_n to denote the returns of a single asset over consecutive time periods, such as consecutive months. Notice that this differs from the notation in Section I.A.2.4, where the X_i were returns on different assets in the same period.

Also, throughout this section we will interpret the X_i as continuously compounded returns, in the sense of (I.A.2.2). With this convention, the return X over the entire period is the sum of the returns over the individual periods,

$$X = X_1 + X_2 + \dots + X_n. \tag{I.A.2.33}$$

To see this, suppose the returns are calculated using (I.A.2.2) from asset prices S_0, S_1, \dots, S_n ; then

$$X_1 + X_2 + \dots + X_n = \ln \frac{S_1}{S_0} + \ln \frac{S_2}{S_1} + \dots + \ln \frac{S_n}{S_{n-1}} = \ln \left(\frac{S_1}{S_0} \frac{S_2}{S_1} \dots \frac{S_n}{S_{n-1}} \right) = \ln \left(\frac{S_n}{S_0} \right).$$

From (I.A.2.33) we see that the effect of serial correlation in a single asset — i.e., correlations between the X_i — can be calculated in the same way as for a portfolio of correlated assets.

Although (I.A.2.33) does not hold exactly for the simple returns in (I.A.2.1), it is sufficiently close that the qualitative conclusions of this section apply to simple returns as well.

Consider the case of uncorrelated returns: the correlation between X_i and X_j is zero, for all $i \neq j$. In this case, it follows from (I.A.2.15) or (I.A.2.26) that

$$\text{Var}[X] = \text{Var}[X_1] + \text{Var}[X_2] + \cdots + \text{Var}[X_n].$$

The return variance over the full horizon is the sum of the variances over the individual periods.

If we further assume that the X_i all have the same variance σ^2 (e.g., all monthly returns have the same variance), we get

$$\text{Var}[X] = n\sigma^2,$$

and then

$$\boxed{\text{StdDev}[X] = \text{StdDev}[X_1 + \cdots + X_n] = \sqrt{n}\sigma.} \quad (\text{I.A.2.34})$$

In the case of $n = 12$ monthly returns, this says that the standard deviation of annual returns is $\sqrt{12}$ times larger than the standard deviation of monthly returns. To annualize a daily standard deviation one would similarly use $\sqrt{250}$ because there are approximately 250 business days in a year.

The fact that the annual standard deviation is only $\sqrt{12}$ times larger than the monthly number and not 12 times larger is similar to the diversification effect we observed in our discussion of portfolio standard deviation. In a portfolio, losses in some assets will tend to be partly offset by gains in others. Similarly, losses in some months will tend to be offset by gains in others.

The “square root of time” formula (I.A.2.34) is often used, for example, to convert a value-at-risk figure from one time horizon to another. However, it is important to stress the two assumptions that underlie this formula:

- the monthly returns have a common standard deviation σ ;
- the monthly returns are uncorrelated with each other.

What happens if we drop the second assumption? From (I.A.2.26) we see that if all the serial correlations ρ_{ij} are positive, then

$$\text{StdDev}[X] = \text{StdDev}[X_1 + \cdots + X_n] > \sqrt{n}\sigma,$$

and the simple rule (I.A.2.34) understates the true variability. Positive serial correlations in returns suggest an asset with a price trend or “momentum”.

If, on the other hand, many of the largest correlations are negative — for example, the consecutive correlations $\rho_{i,i+1}$ — then we may have

$$\text{StdDev}[X] = \text{StdDev}[X_1 + \cdots + X_n] < \sqrt{n}\sigma,$$

in which case (I.A.2.34) overstates the true variability. Negative serial correlations may suggest a mean-reverting price process, in which losses are often followed by gains, and vice versa. Notice that if X_i is negatively correlated with X_{i-1} , and X_{i+1} is negative correlated with X_i , then X_{i+1} and X_{i-1} may be positively correlated with each other. Thus, it may not be possible to have all ρ_{ij} , $j \neq i$, negative.

I.A.2.7 Normally Distributed Returns

None of the ideas discussed thus far in this chapter has relied on any distributional assumptions. In particular, the key identities (I.A.2.14)–(I.A.2.15) relating the mean and variance of $aX + bY$ to those of X , Y and to the correlation between X and Y , do not assume anything about the distribution of X and Y . We now discuss what types of further statements can be made if we assume X and Y are normally distributed.

I.A.2.7.1 The Distribution of Portfolio Returns

Throughout this section, we assume that X and Y are jointly normally distributed. We denote by μ_X , μ_Y , σ_X , σ_Y and ρ_{XY} their means, standard deviations and correlation. As before, a portfolio with weight w on X and $1 - w$ on Y has return

$$\Pi = wX + (1 - w)Y.$$

But any linear combination, say $aX + bY$, of normal random variables is itself normally distributed. Thus, the portfolio return Π is normally distributed if the asset returns are normally distributed.

Furthermore, a normal random variable is characterized by its mean and standard deviation. It follows that if

- we know that the portfolio return Π is normal; and
- we know the mean μ_Π and the standard deviation σ_Π ,

then we know the complete distribution of Π . We know how to find μ_Π and σ_Π from (I.A.2.14) and (I.A.2.15). Knowing the full distribution of Π allows us to answer questions that we could not answer from the mean and standard deviation alone.

I.A.2.7.2 Value-at-Risk

As a first illustration, consider the problem of find a portfolio's value-at-risk (VaR; see also Chapter III.A.1) at a confidence level $1 - \alpha$ (e.g., $\alpha = 0.01$ for a 99% VaR). For this we need to find v such that

$$P(\Pi \leq -v) = \alpha. \tag{I.A.2.35}$$

This is illustrated in the left panel of Figure I.A.2.10, which shows the distribution of Π .

Consider the parameters in (I.A.2.18) and a portfolio with weights $w = 0.6$ and $1 - w = 0.4$. In (I.A.2.19) and (I.A.2.21), we have already calculated

$$\mu_\Pi = 13.2\%, \quad \sigma_\Pi = 21.57\%.$$

These are the mean and standard deviation for the annual return.

Suppose we want to calculate VaR over a 10-day horizon, which is 0.04 years. To convert the annual mean to a 10-day mean we therefore multiply by the number of years, 0.04, to get

$$\mu_\Pi = 0.04 \times 13.2\% = 0.528\%.$$

For the standard deviation, we will use the square root of time rule (I.A.2.34) and thus multiply by $\sqrt{0.04} = 0.2$ to get

$$\sigma_\Pi = 0.2 \times 21.57\% = 4.314\%.$$

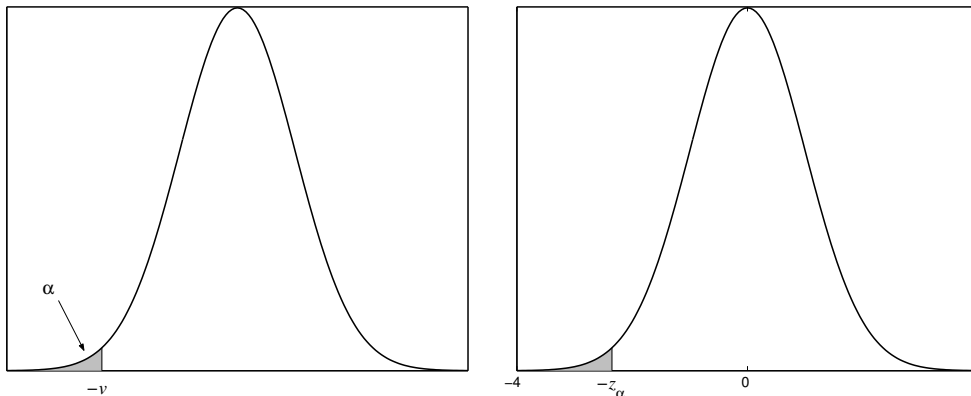


Figure I.A.2.10: Left panel shows the distribution of the portfolio return Π and the value-at-risk v at confidence level $1 - \alpha$. The right panel shows the standard normal distribution and the lower quantile $-z_\alpha$.

Thus, we take the portfolio return Π over a two-week horizon to be normally distributed with mean 0.528% and standard deviation 4.314%.

We need to find v in (I.A.2.35) using this distribution. Because Π is normally distributed, we can represent it as

$$\Pi = \mu_\Pi + Z\sigma_\Pi \quad (\text{I.A.2.36})$$

where Z has a *standard* normal distribution — i.e., one with mean 0 and variance 1. It follows that the point $-v$ can be represented as

$$-v = \mu_\Pi - z_\alpha\sigma_\Pi,$$

where z_α satisfies

$$P(Z \leq -z_\alpha) = \alpha,$$

as illustrated in the right panel of Figure I.A.2.10. In other words, the VaR figure we seek lies z_α standard deviations below the mean.

For $\alpha = 1\%$, we have $z_\alpha = 2.33$. This can be calculated using the Excel function NORMSINV; for example, $\text{NORMSINV}(0.01) = -2.33$. Thus,

$$v = -\mu_\Pi + z_\alpha\sigma_\Pi = -0.528\% + (2.33)(4.314\%) = 9.52\%.$$

Notice that over a two-week horizon, the mean μ_Π is much smaller than the standard deviation σ_Π . In practice, mean returns are also difficult to estimate. It is therefore customary (and usually conservative) to replace the mean over a short horizon with zero. This gives

$$v = z_\alpha\sigma_\Pi = (2.33)(4.314\%) = 10.05\%.$$

VaR is ordinarily reported as a dollar amount, rather than as a percentage. To complete the calculation of VaR, we therefore need to specify the size of the portfolio. Consider a \$5 million portfolio, with \$3 million in XXX and \$2 million in YYY. Over a two-week horizon, its 99% value-at-risk is

$$\$5,000,000 \times 10.05\% = \$502,500.$$

I.A.2.7.3 Probability of Reaching a Target

As our next illustration of the use of the normal distribution in portfolio calculations, we consider the probability of earning \$6 million over one year with a \$10 million investment. We compare three portfolios:

- \$10 million in XXX
- \$10 million in YYY
- \$6 million in XXX and \$4 million in YYY.

In each case, we need to find

$$P(10,000,000\Pi > 6,000,000),$$

where Π is the corresponding portfolio return. Equivalently, we need to find

$$P(\Pi > 0.6).$$

By using the representation (I.A.2.36) in terms of the standard normal random variable Z , we can write this probability as

$$P(\Pi > 0.6) = P\left(\frac{\Pi - \mu_{\Pi}}{\sigma_{\Pi}} > \frac{0.6 - \mu_{\Pi}}{\sigma_{\Pi}}\right) = P\left(Z > \frac{0.6 - \mu_{\Pi}}{\sigma_{\Pi}}\right).$$

For the all-XXX portfolio, $\mu_{\Pi} = \mu_X = 12\%$ and $\sigma_{\Pi} = \sigma_X = 24\%$, so

$$\frac{0.6 - \mu_{\Pi}}{\sigma_{\Pi}} = \frac{0.6 - 0.12}{.24} = 2,$$

and

$$P(Z > 2) = 2.28\%.$$

This can be calculated in Excel as 1-NORMSDIST(2).

For the all-YYY portfolio, $\mu_{\Pi} = \mu_Y = 15\%$ and $\sigma_{\Pi} = \sigma_Y = 30\%$, so

$$\frac{0.6 - \mu_{\Pi}}{\sigma_{\Pi}} = \frac{0.6 - 0.15}{0.30} = 1.5,$$

and

$$P(Z > 1.5) = 6.68\%.$$

For the 60–40 portfolio, we use the values in (I.A.2.19) and (I.A.2.21), which are $\mu_{\Pi} = 13.2\%$ and $\sigma_{\Pi} = 21.57\%$. This gives

$$\frac{0.6 - \mu_{\Pi}}{\sigma_{\Pi}} = \frac{0.6 - 0.132}{0.2157} = 2.17$$

and then

$$P(Z > 2.17) = 1.50\%.$$

The diversified portfolio thus has the *least* chance of reaching the target. The portfolio with the highest risk has the greatest chance of reaching the target. Of course, it also has the greatest chance of experiencing a large negative return. The return distributions of the three portfolios are plotted in Figure I.A.2.11

The fact that the riskiest portfolio maximizes the probability of reaching the target results from the fact that the target of 60% is out in the tail of the return distributions. A trader seeking to maximize the chances of reaching a distant target would similarly choose a high-risk portfolio.

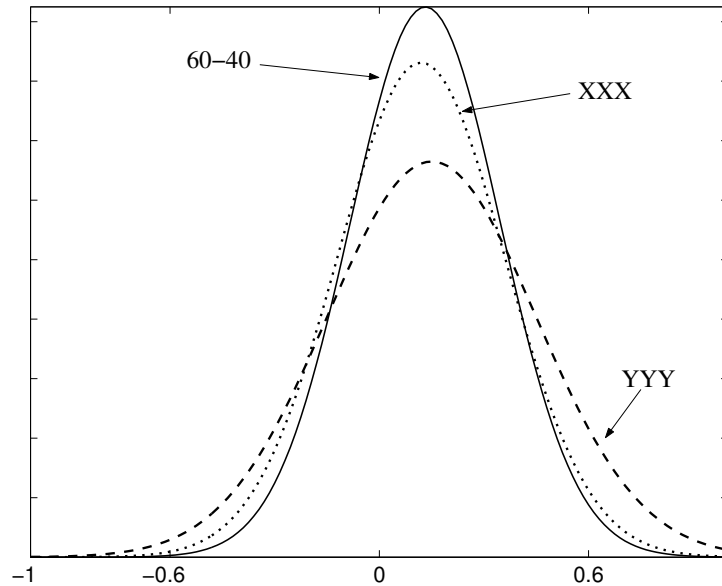


Figure I.A.2.11: Return distributions for three portfolios. The all-YYY portfolio has the greatest chance of exceeding 0.6, but also the greatest chance of falling below -0.6 .

I.A.2.7.4 Probability of Beating a Benchmark

In the previous example, we calculated the probability of exceeding a fixed target. Now we turn to the probability of beating a benchmark which is itself stochastic.

Consider, again, the 60–40 portfolio with a return which we now denote by

$$\Pi_1 = 0.6X + 0.4Y.$$

Suppose this portfolio is to be compared against a benchmark with return

$$\Pi_2 = 0.5X + 0.5Y.$$

What is the probability that the 60–40 portfolio beats the benchmark? In other words, what is

$$P(\Pi_1 > \Pi_2)?$$

To calculate this probability, we rewrite it as

$$P(\Pi_1 - \Pi_2 > 0)$$

which then becomes

$$P(0.6X + 0.4Y - 0.5X - 0.5Y > 0) = P(0.1X - 0.1Y > 0).$$

Because the difference $0.1X - 0.1Y$ is a linear combination of normal random variable, it is itself normally distributed. Its mean is

$$E[0.1X - 0.1Y] = 0.1\mu_X - 0.1\mu_Y = (0.1)(12\%) - (0.1)(15\%) = -0.3\%.$$

Its standard deviation is

$$\text{StdDev}[0.1X - 0.1Y] = \sqrt{0.1^2\sigma_X^2 + 0.1^2\sigma_Y^2 - 2(0.1)(0.1)\sigma_X\sigma_Y\rho_{XY}} = 3.16\%.$$

The probability we seek can thus be written as

$$P(0.1X - 0.1Y > 0) = P\left(\frac{0.1X - 0.1Y + 0.003}{0.0316} > \frac{0 + 0.003}{0.0316}\right) = P(Z > 0.095), \quad (\text{I.A.2.37})$$

with Z a standard normal random variable. Finally,

$$P(Z > 0.095) = 46.2\%.$$

Thus, the 60–40 portfolio has a 46.2% chance of outperforming the benchmark portfolio.

How would this probability change if the assets X and Y were more highly correlated?

If X and Y were more highly correlated, then Π_1 and Π_2 would be more highly correlated. If two portfolio returns are highly correlated, it becomes more difficult for the one with lower mean to outperform the one with higher mean. Thus, we expect that increasing ρ_{XY} should decrease the probability that Π_1 exceeds Π_2 .

That this is indeed the case can be verified from the calculations above. Increasing ρ_{XY} decreases $\text{StdDev}[0.1X - 0.1Y]$ and thus increases the threshold that Z must exceed in (I.A.2.37). This in turn reduces the probability that Z does exceed the threshold.

I.A.3 Capital Allocation

Keith Cuthbertson and Dirk Nitzsche¹

The concepts behind portfolio theory can be used to determine asset allocation strategies. For example, an investor has to decide how to distribute her funds across different equities in different countries. A set of optimal portfolio weights (e.g. 25% of the portfolio's value in UK stocks and 75% of its value in US stocks) can be determined using portfolio theory. An investor might also wish to put some of her own funds in a 'safe asset' (e.g. bank deposit) with the remainder placed in risky equities. Alternatively, if she is not too worried about risk but likes a high return, she may be willing to borrow money at the safe rate (e.g. bank loan), add this to her own funds and put 'the lot' into the stock market. Portfolio theory can also help in this borrowing/lending decision. The concepts used before we 'make it' to the results of mean-variance portfolio theory are quite numerous and somewhat complex. Hence, it is useful at the outset to sketch the main concepts and ideas we will meet and draw out some basic implications of the approach.

I.A.3.1 An Overview

We restrict our world to one in which investors can choose a set of risky assets (stocks) plus an asset that is risk-free over the fixed holding period (e.g. fixed-term bank deposit or a Treasury bill). Investors can borrow and lend as much as they like at the risk-free rate. We assume investors like higher expected returns but dislike risk (i.e. they are risk-averse). The expected return on an *individual* security we denote ER_i , and we assume that the risk on an *individual* security i can be measured by the variance σ_i^2 or standard deviation σ_i of its return. We assume all individuals form the same (i.e. homogeneous) expectations about expected returns and the variances and covariances (correlation) between the various returns. Transaction costs and taxes are assumed to be zero.

I.A.3.1.1 Portfolio Diversification

σ_{12} is the covariance between two returns and is related to the correlation coefficient ρ by the formula:

$$\rho = \sigma_{12} / \sigma_1 \sigma_2 \quad (\text{I.A.3.1})$$

¹Keith Cuthbertson is Professor of Finance and Dirk Nitzsche is Senior Lecturer at the Cass Business School, City University, London.

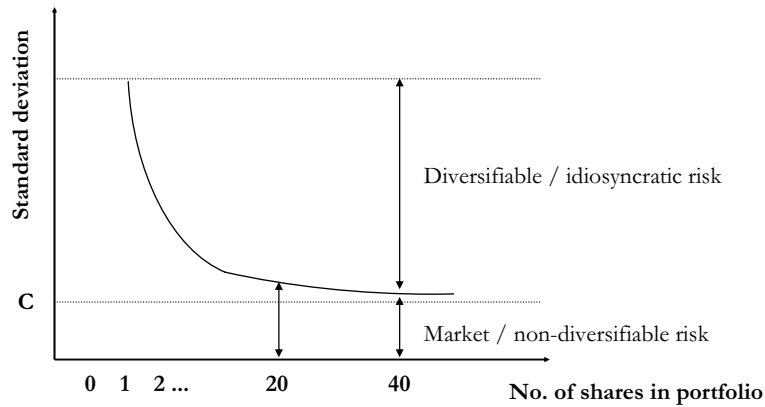
The covariance and correlation coefficient will both have the same sign, but the covariance has the annoying property that it is dependent on the units used to measure returns (e.g. proportions or percentages), whereas the correlation coefficient is ‘dimensionless’ and must always lie between +1 and –1. In the extreme case where $\rho = +1$, the two assets’ returns are perfectly positively (linearly) related and therefore the asset returns *always* move in the same *direction* (but not necessarily by the same percentage amount). For $\rho = -1$ the converse applies and for $\rho = 0$ the asset returns are not (linearly) related. As we see below, the ‘riskiness’ of the portfolio consisting of both asset 1 and asset 2 depends crucially on the sign and size of ρ . If $\rho = -1$, risk may be completely eliminated by holding a specific proportion of your wealth in both assets. Even if ρ is positive (but less than +1) the riskiness of the overall portfolio is reduced (although not to zero).

Consider the reason for holding a *diversified portfolio* consisting of a *set* of risky assets. Assume for the moment that you use all your ‘own funds’ to invest in stocks. Putting all your wealth in asset 1, your forecast or ‘expected’ return is ER_1 with risk of σ_1^2 . Similarly, holding just asset 2 you expect to earn ER_2 and incur risk σ_2^2 . Let us assume a two-asset world where there is a negative covariance of returns $\sigma_{12} < 0$. (This also implies a negative correlation coefficient $\rho_{12} = \sigma_{12} / \sigma_1 \sigma_2$.) When the return on asset 1 rises, that on asset 2 tends to fall, so if you hold both assets the risks are partially offsetting. Hence, if you diversify and hold both assets, this would seem to reduce the variance of the *overall* portfolio return (i.e. of asset 1 plus asset 2). To simplify even further, suppose that $ER_1 = ER_2 = 10\%$ and $\sigma_1^2 = \sigma_2^2$. In addition, assume that when the return on asset 1 increases by 1%, that on asset 2 falls by 1% (i.e. returns are perfectly negatively correlated, $\rho = -1$). Under these conditions when you hold half your initial wealth in each of the risky assets, the expected return on the overall portfolio is $ER_p = 0.5 ER_1 + 0.5 ER_2 = 10\%$. However, diversification has reduced the risk on this portfolio to zero; an above average return on asset 1 is always matched by an equal below average return on asset 2 (since $\rho = -1$). Our example is a special case. But, in general, even if the correlation between returns is zero or positive (but not perfectly positively correlated), it still pays to diversify and hold a combination of both assets.

So, the benefits of diversification in reducing risk depend on returns having less than perfect (positive) correlation. In fact, even a little diversification quickly reduces risk. Suppose we *randomly* choose a one-stock portfolio, a two-stock portfolio, ... , n -stock portfolio from stocks in the S&P 500 index, by throwing the required number of darts at the stocks page of the *Wall Street Journal*. Each time we throw the darts, we calculate the *portfolio* standard deviation σ_p (with weights $w_i = 1/n$). We find that ‘risk’ drops to a level C with only about 20–30 stocks (Figure I.A.3.1).

This is because the risk that is *specific* to each firm/industry is random around zero (e.g. risk due to strikes, unfilled orders, managerial incompetence, the weather, computer failures, production line malfunctions, etc.). When averaged across *many* firms (stocks), these specific risks cancel each other out (i.e. good and bad luck) and hence do not contribute to overall *portfolio risk*. Hence, we can eliminate *specific* (or *diversifiable* or *idiosyncratic*) risk, simply by adding more stocks to our portfolio. Hence, in general, portfolio diversification arises from $\rho < +1$, and the law of large numbers, $n \rightarrow \infty$.

Figure I.A.3.1: Increasing size of portfolio



For example, consider $n = 2$ for illustrative purposes. If the proportion of our own wealth held in asset 1 is w_1 , then portfolio variance is:

$$\sigma_p^2 = w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2w_1 w_2 \rho \sigma_1 \sigma_2 \quad (w_1 + w_2 = 1) \quad (\text{I.A.3.2})$$

- if $\rho = +1$, then $\sigma_p = w_1 \sigma_1 + w_2 \sigma_2$ and σ_p is a (linear) weighted average of σ_1 and σ_2 and there are no diversification benefits;
- if $\rho < +1$, then $\sigma_p < w_1 \sigma_1 + w_2 \sigma_2$ and σ_p must be less than the weighted average of σ_1 and σ_2 . This is the portfolio diversification effect.

If $\rho = 0$, equation (I.A.3.2) becomes:

$$\sigma_p^2 = w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 \quad (\text{I.A.3.3})$$

For example, suppose $\sigma_1 = \sigma_2 = \sigma$ and $w_1 = 1/2$, then $\sigma_p = \sigma / \sqrt{2}$. In general, as we see below, for $\rho = 0$ we have $\sigma_p = \sigma / \sqrt{n}$. This is the ‘law of large numbers’ or ‘insurance effect’. A large number of *uncorrelated* ($\rho = 0$) events has a low variance (i.e. $\sigma_p \rightarrow 0$ as $n \rightarrow \infty$). This is why your car insurance premium is low relative to the replacement value of the car. If the car insurer has a large number of customers who have accidents that are largely independent of each other, then the risk of the whole ‘portfolio’ of customers is relatively small. (The reason, in practice, why the risk does not actually reach zero as the number of customers increases is that there is some small positive correlation between accidents from claimants within the same company, e.g. on the relatively few days when road conditions are particularly bad throughout the entire country in which the insurer operates.)

The *systematic risk* of a portfolio is defined as the risk that cannot be diversified away by adding extra securities to the portfolio. (It is also referred to as ‘non-diversifiable’, ‘irreducible’, ‘portfolio’ or ‘market’ risk.) There is always some non-zero risk even in a well-diversified portfolio, and this is because different firms are affected by economy-wide factors (e.g. changes in interest rates, exchange rates, tax laws). This is what gives rise to the correlation between different stock returns – but the correlation is not perfect, because the economy-wide variables affect different firms (profits) by differing amounts. To see the influence of these correlated ‘events’ on *portfolio* variance consider a portfolio of n assets held in proportions w_i ($0 < w_i < 1$):

$$\sigma_p^2 = \sum_{i=1}^n w_i^2 \sigma_i^2 + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n w_i w_j \sigma_{ij} \quad (\text{I.A.3.4})$$

With n assets there are n variance terms σ_i^2 and $n(n-1)/2$ covariance terms σ_{ij} that contribute to the variance of the portfolio. The number of covariance terms rises much faster than the number of assets in the portfolio. To illustrate the dependence of σ_p^2 on the covariance terms, consider a simplified portfolio where all assets are held in the same proportion ($w_i = 1/n$) and where all variances and covariances are *constants* (i.e. $\sigma_i^2 = V$ and $\sigma_{ij} = C$). Then equation (I.A.3.4) becomes:

$$\sigma_p^2 = n \left(\frac{1}{n^2} V \right) + n(n-1) \left(\frac{1}{n^2} C \right) = (1/n)V + (1-1/n)C \quad (\text{I.A.3.5})$$

It follows that as $n \rightarrow \infty$ the influence of the variance term V approaches zero and σ_p^2 equals the (constant) covariance, C . Thus the variance of *individual* securities that represents (idiosyncratic or specific) risk particular to that firm or industry can be diversified away – you might surmise at

this point that if you hold a diversified portfolio then you should not earn any return from holding the specific risk of these firms. When we discuss the CAPM in Chapter I.A.4, we do indeed find that this holds true. However, covariance risk cannot be diversified away, and it is the covariance terms that (in a loose sense) give rise to market/non-diversifiable or systematic risk (Figure I.A.3.1). As we shall see in Chapter I.A.4, this ‘covariance’ or market risk can be represented by the beta of the security.

Above we have shown that, in general, the investor can reduce portfolio risk σ_p by including additional stocks in her portfolio. In fact the portfolio variance σ_p^2 falls very quickly as one increases the number of stocks held from 1 to around 25 and thereafter the reduction in portfolio variance is quite small (Figure I.A.3.1). This, coupled with the brokerage fees and information costs of monitoring a large number of stocks, may explain why individuals tend to invest in only a relatively small number of stocks. Individuals may also obtain the benefits of diversification by investing in mutual funds (i.e. unit trusts), ‘closed-end’ mutual funds (i.e. investment trusts) and pension funds, since these take funds from a large number of individuals to invest in a wide range of financial assets and each individual then owns a proportion of this ‘large portfolio’.

I.A.3.1.2 Tastes and Preferences for Risk versus Return

The reader may now be surmising that the individual investor’s tastes or preferences must come into the analysis at some point, and that would be correct. However, there is a quite remarkable result, known as the *separation principle*. The investment decision can be broken down into two separate decisions. The first decision concerns the choice of the *optimal proportions* w_i^* of risky assets held, and this is *independent of the individual’s preferences* concerning her subjective trade-off between risk and return. This choice depends only on the individual’s views about the objective market variables, namely expected returns, variances and covariances. Hence, we can use portfolio theory to determine the proportions to place in the *risky* assets. Now if we are also willing to assume that expectations about variances etc. are the same for all investors (and it’s a big ‘if’), then we find that all investors hold the *same* proportions of the risky assets (e.g. all investors hold 1/20 of ‘A-shares’, 1/80 of ‘B-shares’, etc.), irrespective of their preferences. Hence aggregating, all individuals will hold these risky asset in the same proportions as in the (aggregate) market portfolio (e.g. if Microsoft shares represent 5% of the total stock market index by value, then all investors hold 5% of their *own* risky asset portfolio in Microsoft shares). Note that even if all investors do not have the same view about variances, expected return, etc., then portfolio theory can still be used to calculate a *particular investor’s* optimal asset allocation strategy. So all is not lost! However, we should not call the resulting portfolio ‘the market portfolio’.

It is only after mimicking the market portfolio that investors’ preferences enter the calculation. In the second stage of the decision process the investor decides how much to borrow (lend) in

order to augment (reduce) the amount of her own initial wealth invested, in the fixed proportions, in the market portfolio of risky assets. Suppose you have a *very* risk-averse investor. Such an investor, faced with the choice between (i) a certain gain of \$5 and (ii) a 50–50 chance of losing \$10 or gaining \$10,000, would choose option (i). Most people are not this risk-averse, but a very risk-averse investor will put most of her own wealth into the risk-free asset (which pays r) and will invest only a small amount of her own wealth in the risky assets, in the fixed proportions w_i^* . The converse applies to a much less risk-averse person, who will *borrow* at the risk-free rate and use these proceeds (as well as her own initial wealth) to invest in the fixed bundle of risky assets in the optimal proportions w_i^* . Note however, that the *proportion* w_i^* of the total funds in the risky assets will be the same for both the very risk-averse and less risk-averse person (but the *dollar* amounts will be different, as we shall see). Now, let us examine the ‘building blocks’ of portfolio theory in more detail.

I.A.3.2 Mean–Variance Criterion

Throughout this chapter we shall use the following equivalent ways of expressing expected returns, variances and covariances:

- expected return $\equiv \mu_i \equiv ER_i$
- variance of returns $\equiv \sigma_i^2 \equiv \text{var}(R_i)$
- covariance of returns $\equiv \sigma_{ij} \equiv \text{cov}(R_i, R_j)$

We assume that the investor would prefer a higher expected *portfolio* return ER_p to a lower expected return, but she dislikes risk (i.e. is risk-averse). We choose to measure risk by the variance (or standard deviation) of the returns on the *portfolio* of risky assets. More risk implies less ‘satisfaction’. In the jargon of economics, the investors’ utility (U) is assumed to depend only on expected return and the variance (or standard deviation) of the return:

$$U = U(ER_p, \sigma_p) \quad (\text{I.A.3.6})$$

Thus, if the agent is presented with a portfolio A (of n securities) and a portfolio B (of a different set of n securities), then, according to the mean–variance criterion, portfolio A is preferred to portfolio- B if

$$E_A(R_p) \geq E_B(R_p) \quad (\text{I.A.3.7a})$$

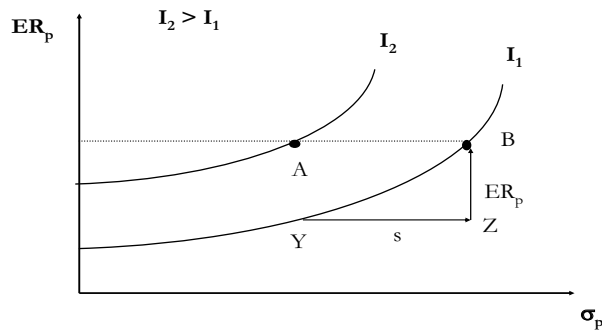
and

$$\sigma_A^2(R_p) \leq \sigma_B^2(R_p) \quad \text{or equivalently} \quad \sigma_A(R_p) \leq \sigma_B(R_p) \quad (\text{I.A.3.7b})$$

where σ_i ($i = A, B$) is the standard deviation of the return of the portfolio.

Of course, if, for example, $E_A(R_p) > E_B(R_p)$ but $\sigma_A(R_p) > \sigma_B(R_p)$, then we cannot say which portfolio the investor prefers. Portfolios that satisfy the mean–variance criterion are known as the set of *efficient portfolios*. A portfolio A that has a lower expected return *and* a higher variance than another portfolio B is said to be ‘inefficient’ with respect to portfolio B and an individual would (in principle) never hold portfolio A (i.e. portfolio A is ‘dominated’ by portfolio B).

Figure I.A.3.2: Individual preferences



We can represent any individual’s preferences in an (ER_p, σ_p) graph (Figure I.A.3.2). At all points on the indifference curve I_1 , the individual is equally ‘happy’ (i.e. has constant utility). If you increase risk from Y to Z , then in order to compensate for this loss of satisfaction the individual must be given a higher expected return (Z to B) to return to the same level of happiness. Indifference curve I_2 represents higher levels of satisfaction than I_1 . The individual would prefer to be at A rather than B since both points have the same expected return but at A the individual bears less risk.

I.A.3.3 Efficient Frontier: Two Risky Assets

To simplify matters we assume there are only four possible scenarios (see Table I.A.3.1). A ‘high’ level of interest rates is detrimental to equity returns but high (real) growth in the economy leads to high expected returns. This is because high interest rates (tight monetary policy) generally imply lower profits while high growth implies high profits. Some of these profits will then be

distributed as dividends and may lead to capital gains on the stock. The four combinations of possible scenarios have an equal probability of occurrence ($1/4$ each).

Table I.A.3.1: Two risky assets

State	Interest	Growth	Probability	Return	
				Equity 1	Equity 2
1	High	Low	0.25	-5%	45%
2	High	High	0.25	5%	35%
3	Low	Low	0.25	10%	10%
4	Low	High	0.25	25%	-5%

Given the returns (R_1, R_2) on the two assets *under each scenario*, in Table I.A.3.1 we can calculate the expected return.

$$ER_1 = -5\%/4 + 5\%/4 + 10\%/4 + 25\%/4 = 8.75\% \quad (\text{I.A.3.8a})$$

and similarly:

$$ER_2 = 21.25\% \quad (\text{I.A.3.8b})$$

Note that the expected return, using probabilities, is the same as the *sample average* of the return, if we had observed $R_1 = (-5, 5, 10, 25)$ over a four-year period. The variance and standard deviation are given by

$$\sigma_1^2 = \frac{1}{4} [(-5\% - 8.75\%)^2 + (5\% - 8.75\%)^2 + (10\% - 8.75\%)^2 + (25\% - 8.75\%)^2] = 117.28$$

(I.A.3.9a)

so $\sigma_1 = \sqrt{117.28} = 10.83$, and similarly

$$\sigma_2^2 = 392 \quad (\text{I.A.3.9b})$$

and $\sigma_2 = \sqrt{392} = 19.8$. We can also work out the covariance and correlation coefficient between the returns on these two assets. The returns on the two assets are negatively correlated. When the return on asset 1 is high, then that on asset 2 tends to be low. The covariance and correlation coefficient are

$$\begin{aligned} \sigma_{12} &= [(-5-8.75)(45-21.25)]/4 + [(5-8.75)(35-21.25)]/4 \\ &+ [(10-8.75)(10-21.25)]/4 + [(25-8.75)(-5-21.25)]/4 = -204.68 \end{aligned} \quad (\text{I.A.3.10})$$

$$\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{-204.68}{10.83 \times 19.8} = -0.9549 \quad (\text{I.A.3.11})$$

These results are summarised in Table I.A.3.2.

Table I.A.3.2: Summary statistics (two risky assets)

	Risky Assets	
	Equity 1	Equity 2
Mean, ER_I	8.75%	21.25%
Std. dev., σ_I	10.83%	19.80%
Correlation(Equity 1, Equity 2)		−0.9549
Covariance(Equity 1, Equity 2)		−204.688

Suppose the investor at this stage is not allowed to borrow, or lend, the safe asset. What opportunities are open to her when faced with assets 1 and 2 with variances σ_1^2 and σ_2^2 and covariance σ_{12} or correlation ρ between the two returns? We can calculate the combinations of expected return and standard deviation of return available to her as she varies the proportion of her *own* wealth (remember, there is no borrowing or lending at this stage) held in either equity 1 or equity 2. We begin with an algebraic exposition but then demonstrate the points made using a simple numerical example. Suppose the investor chooses to hold a proportion of her wealth (w_1) in asset 1 and a proportion $w_2 = (1 - w_1)$ in asset 2. The *actual return* on this diversified portfolio (which will not be revealed until one period later) is:

$$R_p = w_1R_1 + w_2R_2 \quad (\text{I.A.3.12a})$$

The *expected* return on the portfolio is:

$$ER_p = w_1ER_1 + w_2ER_2 \quad (\text{I.A.3.12b})$$

The *variance of the portfolio* is given by:

$$\sigma_p^2 = w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1w_2(\rho\sigma_1\sigma_2) \quad (\text{I.A.3.13})$$

where we have used $\sigma_{12} = \rho\sigma_1\sigma_2$. The question we now ask is: how do ER_p and σ_p vary, relative to each other, as the investor alters the proportion of her own wealth held in each of the risky assets? This is the *feasible set* or *opportunity set*.

Remember that ER_1 , ER_2 , σ_1 , σ_2 and σ_{12} (or ρ) are fixed and known and we simply alter the proportions of wealth: w_1 in asset 1 and $w_2 (= 1 - w_1)$ in asset 2. (Note that there is no maximisation/minimisation problem here. It is a purely *arithmetic* calculation given the definitions of ER_p and σ_p .) A numerical example is given in Table I.A.3.3.

Table I.A.3.3: Portfolio risk and return

State	Share of		Portfolio	
	Equity 1	Equity 2	ER_p	σ_p
	w_1	w_2		
1	1	0	8.75%	10.83%
2	0.75	0.25	11.88%	3.70%
3	0.5	0.5	15%	5%
4	0	1	21.25%	19.80%

For example, for $w_1 = 0.75$, $w_2 = 0.25$ we get:

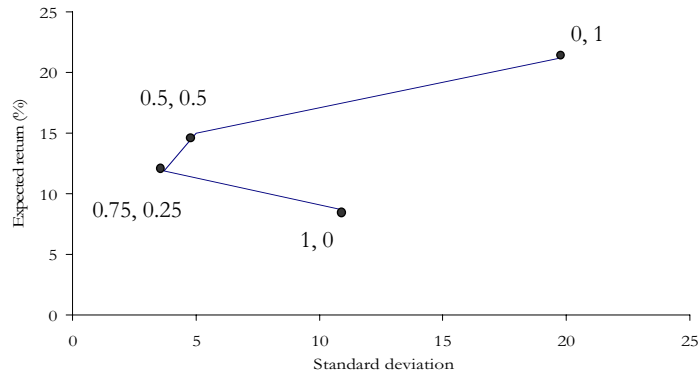
$$ER_p = 0.75 \times 8.75\% + 0.25 \times 21.25\% = 11.88\% \quad (\text{I.A.3.14})$$

$$\sigma_p^2 = 13.7 \quad (\text{I.A.3.15})$$

so $\sigma_p = 3.7$. This is noted in Figure I.A.3.3. The risk–return combinations in Figure I.A.3.3 represent an *opportunity set* for every investor. However, the investor would never choose points along the lower portion of the curve because points along the upper portion have a higher expected return but no more risk, hence the locus of points above and to the right of (0.75, 0.25) is known as the efficient frontier.

For two risky assets, the efficient set is a ‘curved’ locus of points in (ER_p, σ_p) space, each point representing different proportions (w_1 and w_2) of the two risky assets. It is clear that combining the two assets into a portfolio gives the investor a wider set of risk–return combinations than holding either one of the two assets. This is the principle of diversification.

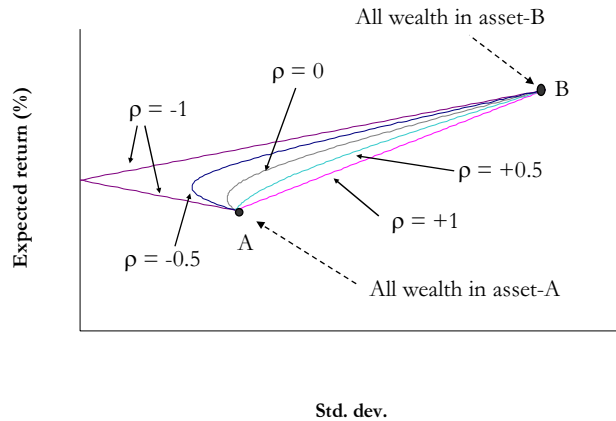
Figure I.A.3.3: Efficient frontier



I.A.3.3.1 Different Values of the Correlation Coefficient

At any point in time there will be only one value for the correlation coefficient ρ , given by the (past) behaviour of the two returns. However, it is interesting to see what happens to the efficient frontier when ρ moves from +1 to -1. (In the real world ρ may change over time, so later in time we may have a ‘new’ efficient frontier.) This is done in Figure I.A.3.4, where we construct the mean–variance combinations for different values of ρ . In general as ρ approaches -1 the (μ_p, σ_p) locus moves closer to the vertical axis, indicating that a greater reduction in portfolio risk is possible for any given expected return. For $\rho = -1$, the curve hits the vertical axis, indicating there is some value for the proportions (w_1, w_2) held in the two assets that reduces risk to zero. For $\rho = 1$ the risk–return locus is a straight line between the (μ_i, σ_i) points for each individual security – there is no diversification effect along the straight line AB.

Figure I.A.3.4: Altering the correlation coefficient



I.A.3.4 Asset Allocation

If an investor wants to minimise the risk of her portfolio our analysis would indicate that she should hold the risky assets in the proportions (0.75, 0.25) (Figure I.A.3.3). On the other hand, if the investor indicates she is willing to tolerate risk of only $\sigma = 5\%$, then the portfolio manager would suggest that the highest return she could *expect* to obtain was 15%, and this would require an asset allocation of (0.5, 0.5). The converse of this argument is that if the investor asks the portfolio manager how much risk she would have to bear to achieve an expected return of, say, 20%, the answer would be $\sigma = 19.8\%$ with asset allocation (0, 1). All of the above assumes no borrowing or lending but only investing the investor’s own wealth. Hence, with no borrowing or lending allowed, the investor’s allocation (i.e. proportions) held in the two risky assets *does* depend on their risk tolerance level. However, somewhat counter intuitively, once we allow borrowing and lending, the client’s optimal *proportions* in the risky assets are independent of her risk tolerance.

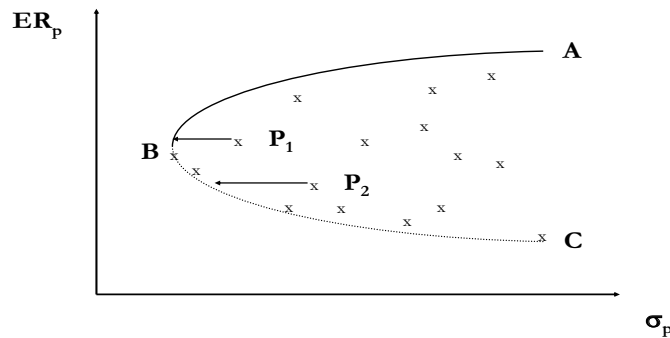
I.A.3.4.1 The efficient frontier: n risky assets

We now consider the case of n assets. When we vary the proportions w_i ($i = 1, 2, \dots, n$) to form portfolios, it is obvious that there is potentially a large number of such portfolios. We can form asset portfolios consisting of 2, 3, ... or n assets. We can also form portfolios consisting of a fixed number of assets but in different proportions. All of these possible portfolios are represented by the points on and inside the convex ‘egg’ (Figure I.A.3.5). These points represent the opportunity set of portfolio weights (the w_i) of the risky assets.

If we apply the mean–variance criterion, then all of the points in the interior of the *portfolio opportunity set* (e.g. P_1, P_2 in Figure I.A.3.5) are (mean–variance) dominated by those on the curve ABC since the latter have a lower variance for a given expected return. Points on the curve AB also dominate those on BC, so the curve AB represents the proportions w_i in the *efficient set* of portfolios and again is referred to as the *efficient frontier*.

- each point on the efficient frontier represents one risky-asset bundle.
- each bundle comprises n risky assets held in the fixed proportions w_i .

Figure I.A.3.5: Efficient frontier (=AB)



Point B (Figure I.A.3.5) represents a particular mix of risky assets, which gives rise to the (*global*) *minimum variance portfolio* – this is the lowest possible variance you can obtain by combining the risky assets. Note that the (*global*) minimum variance point still gives a positive portfolio expected return, since although at this point you are subject to minimum risk possible, nevertheless you are still holding some risk and therefore should be rewarded by a positive expected return. In our mean–variance approach you will not choose this minimum variance point because you are willing to hold both more risk and more return and therefore you end up to the north-east of point B.

I.A.3.5 Combining the Risk-Free Asset with Risky Assets

Let us now take *one* risky bundle (of n assets already held in *fixed* proportions w_i). For $n = 3$ we might have $w_1 = 20\%$, $w_2 = 25\%$ and $w_3 = 55\%$, which makes up our one risky bundle. Now allow investors to borrow or lend at the safe rate of interest r . Because r is fixed over the holding

period, the variance on the risk-free asset is zero, as is its covariance with our own risky bundle.

Thus, the investor can:

- invest all of her wealth in the one risky bundle and undertake no lending or borrowing;
- invest less than her total wealth in the risky bundle and use the remainder to lend at the risk-free rate;

or

- invest more than her total wealth in the risky bundle by borrowing additional funds at the risk-free rate; in this case she is said to hold a *levered portfolio*.

The above choices are represented in the *transformation line*, which is a relationship between expected return and risk for a portfolio consisting of one safe asset plus one risky bundle. The transformation line holds for *any* portfolio consisting of these two assets and it turns out that the relationship between expected return and risk (measured by the standard deviation of the ‘new’ portfolio) is a straight line.

To derive a particular transformation line, consider the risk-free return r (on, say, a T-bill) and the return on a single ‘bundle’ of risky assets R_q (Table I.A.3.4). Since $r = 10\%$ for all scenarios, then $\sigma_r = 0$. The risky-asset bundle q we assume has a mean return $ER_q = 22.5\%$ and a standard deviation $\sigma_q = 24.87\%$. The expected return on this new ‘two-asset’ portfolio is

$$ER_N = \varkappa r + (1 - \varkappa)ER_q \tag{I.A.3.16}$$

Table I.A.3.4: Risk-free and risky ‘bundle’

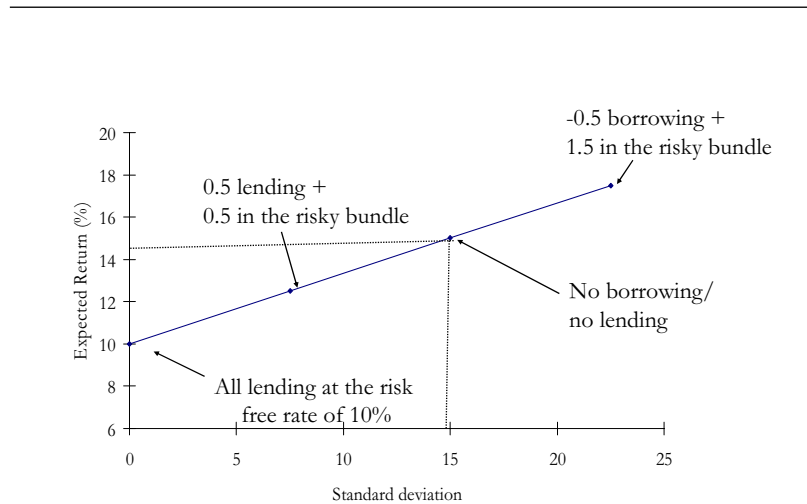
	Returns	
	T-bill (safe)	Equity (risky)
Mean	$r = 10\%$	$r_q = 22.5\%$
Std. dev.	$\sigma_r = 0$	$\sigma_q = 24.87\%$

We can now alter the proportions \varkappa held in the risk-free asset and $1 - \varkappa$ held in the risky bundle to obtain a ‘new’ portfolio. The possible combinations of expected return ER_N and risk σ_N on this ‘new’ portfolio are shown in the last two columns of Table I.A.3.5. This *linear opportunity set* is shown in Figure I.A.3.6 as the transformation line.

Table I.A.3.5: ‘New’ portfolio: (transformation line)

State	Share of Wealth in		‘New’ Portfolio	
	T-bill	Equity	ER_N	σ_N
	w_1	w_2		
1	1	0	10%	0%
2	0.5	0.5	16.25%	12.44%
3	0	1	22.5%	24.87%
4	-0.5	1.5	28.75%	37.31%

Figure I.A.3.6: Transformation line: one riskless asset and one risky ‘bundle’



The transformation line has an intercept equal to the risk-free rate ($r = 10\%$). Here the investor puts *all* her wealth in the safe asset ($x = 1$). When *all* of the investor’s own wealth is held in the risky bundle ($x = 0$) then the ‘new’ portfolio has $ER_N = 22.5\%$ and $\sigma_N = 24.87\%$. These are of course the expected return and standard deviation of the risky-asset bundle (ER_q and σ_q). This is the ‘no borrow/no lend’ portfolio.

When $x = 1$, all wealth is invested in the risk-free asset and $ER_N = r$ and $\sigma_N = 0$. For $0 < x < 1$ some wealth is invested in the risk-free asset and the remainder is put in the risky-asset bundle. When $x = 0$, all the investor’s wealth is invested in stocks and $ER_N = ER_q$. For $x < 0$ the agent borrows money at the risk-free rate r to invest in the risky asset. For example, when $x = -0.5$ and initial wealth is \$100, the individual borrows \$50 (at an interest rate r) and invests \$150 in stocks

(i.e. a levered position). From Table I.A.3.5 we see that this gives $ER_N = 28.75\%$ and $\sigma_N = 37.31\%$ (see also Figure I.A.3.6).

The transformation line gives us the *linear* risk–return relationship for *any* portfolio consisting of a combination of investment in the safe asset *and* one ‘bundle’ of risky assets. At each point on any transformation line the investor holds the risky assets in the *same fixed proportions* w_i .

All the points (except the intercept) on the transformation line represent *fixed* proportions $w_i = 20\%$, 25% and 55% (say) in our one risky bundle (of $n = 3$ risky assets): ‘alpha’, ‘beta’ and ‘gamma’. The only ‘quantity’ that varies along the transformation line is the proportion held in the *one* risky bundle of assets relative to that held in the *one* safe asset. The investor can borrow or lend and be anywhere along the transformation line. For example, the (0.5, 0.5) point (Figure I.A.3.6) represents 50% in the safe asset and 50% in the single bundle of risky securities. Hence, an investor with \$100 would hold \$50 in the risk-free asset and \$50 in the one risky bundle made up of $0.2 \times \$50 = \10 in alpha, $0.25 \times \$50 = \12.50 in beta and $0.55 \times \$50 = \27.50 in the gamma securities.

Since r is known and fixed over the holding period, the standard deviation of the ‘new’ portfolio depends only on the standard deviation of the one risky bundle and this is why the opportunity set in this case is a straight line. For *any* portfolio consisting of two assets, one of which is a single risky bundle and the other is a safe asset, the relationship between the expected return on this new portfolio ER_N and its standard deviation σ_N is *linear* with intercept r . When a portfolio consists only of n *risky* assets then, as we have seen, the efficient frontier in (ER_p, σ_p) space is curved. This should not be unduly confusing since the portfolios considered in the two cases are very different.

If we choose a single risky bundle with $\sigma_k = 30\%$ (and $ER_k = 15\%$) then we can draw the corresponding transformation line L (Figure I.A.3.7). Similarly for our original single risky bundle $\sigma_q = 24.87\%$ ($ER_q = 22.5\%$) we have a higher transformation line L'. Hence, each single risky bundle (each with *different but fixed* weights w_i) has its own transformation line.

Figure I.A.3.7: Transformation lines

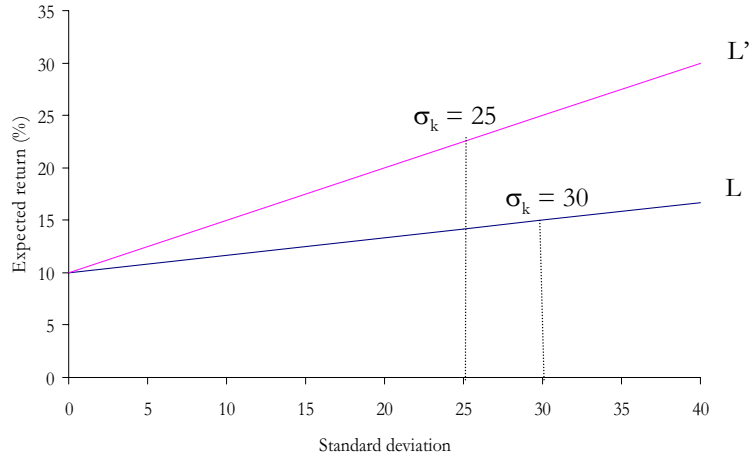
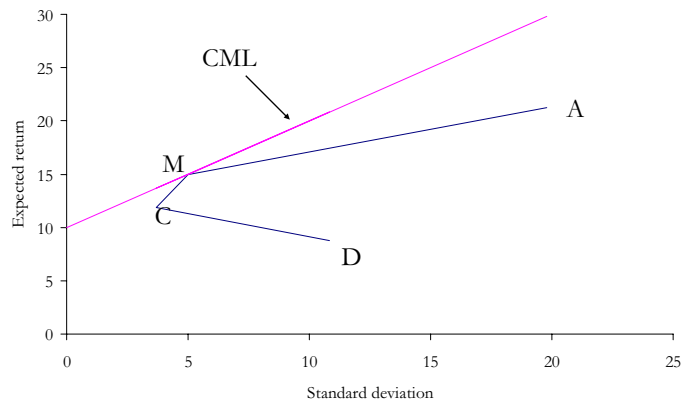


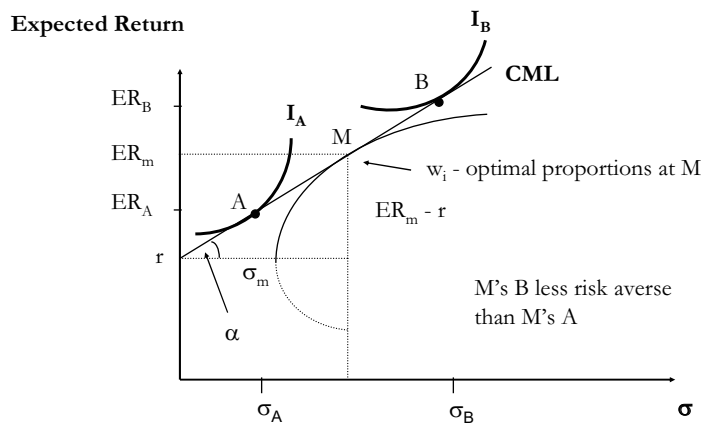
Figure I.A.3.8: Efficient frontier and CML



I.A.3.6 The Market Portfolio and the CML

Although an investor can attain any point along L in Figure I.A.3.7, *any* investor (regardless of her preferences) would prefer to be on the transformation line L'. This is because at any point on L' the investor has a greater expected return for any given level of risk compared with points on L. The 'highest' transformation line possible is the one that is tangent to the efficient frontier and it provides the investor with the highest possible return per unit of risk. Point M in Figure I.A.3.8 would generally represent a 'bundle' of n risky assets held in certain *fixed* proportions (w_i). In our simple model with only two risky assets these optimal proportions at M are seen to be $w_1 = 0.5$ and $w_2 = 0.5$. M is always a single bundle of stocks held in fixed proportions, by *all* investors. Hence, point M is known as the *market portfolio* and L' is known as the *capital market line* (CML).

Figure I.A.3.9: CML and market portfolio



The CML is the transformation line that is 'tangent' to the efficient frontier. When we have many risky assets the efficient frontier is a smooth curve (Figure I.A.3.9) and the market portfolio M represents the optimal proportions (w_i^*) of the risky assets held by all investors. At M the investor does not borrow or lend. Notice that M represents that point on the efficient frontier that maximises

$$\tan \theta = \frac{ER_m - r}{\sigma_m} \quad (\text{I.A.3.17})$$

The optimal proportions w_i maximise expected (excess) return $ER_m - r$ per unit of *portfolio* risk σ_p . In fact mathematically $\tan \theta$ is a nonlinear function of the w_i (see equations (I.A.3.12b) and (I.A.3.13)) and maximising $\tan \theta$ gives the optimal values for the w_i at point M (see Appendix).

I.A.3.7 The Market Price of Risk and the Sharpe Ratio

If all individuals are at the point represented by M, then they all have the *same* excess return per unit of risk. The *market price of risk* is given by:

$$\lambda_m = \frac{ER_m - r}{\sigma_m} \quad (\text{I.A.3.18})$$

If every investor had homogeneous expectations, then point M and hence λ_m would be the same for all investors. However, if investors have heterogeneous expectations then this ratio can be different for each investor and this allows us to use the Sharpe ratio as a measure of the relative performance for different investors. If an investor holds *any portfolio p* of stocks then the Sharpe ratio is defined as:

$$S = (ER_p - r) / \sigma_p \quad (\text{I.A.3.19})$$

The higher is S , the higher is the expected (excess) return $ER_p - r$ per unit of portfolio risk, σ_p . Thus for two investment managers, the one whose portfolio has a higher Sharpe ratio may be deemed the more successful in trading off return against risk. The Sharpe ratio is therefore a type of *performance index* that can be used to rank the relative success of different investment managers or investment strategies. In practice, ER_p , σ_p and r are measured by their sample averages over the whole ‘performance period’ under study (e.g. based on the past 60 months of returns data).

I.A.3.8 Separation Principle

If we now allow investors to borrow or lend, they will ‘mix’ the risk-free asset with the risky bundle represented by M, to get to their *preferred* position along the CML. Hence, investors’ preferences determine at which point along the CML each *individual* investor ends up. For example an investor with little or no risk aversion would end up at a point like B (Figure I.A.3.9), where she borrows money (at r) to augment her own wealth and then invests the borrowed money and all her own wealth in the risky bundle represented by M (but she still holds *all* her n risky assets in the fixed *proportions* w_i^*).

The investor makes two separate decisions:

1. Knowledge of expected returns, variances and covariances determines the efficient frontier. The investor then determines point M as the *point of tangency* of the line from r to the efficient frontier. All this is accomplished without any recourse to the individual's preferences. All investors, regardless of preferences (but with the same view about expected returns, etc.), will 'home in' on the portfolio proportions (w_i^*) of the *risky* securities represented by M. All investors hold the *market portfolio*, or more correctly, all investors hold their n risky assets in the same proportions as their relative value in the market. Thus, if the value of Microsoft shares constitutes 10% of the total stock market valuation, each investor holds 10% of her own risky portfolio in Microsoft shares.

2. The investor now determines how she will combine the market portfolio (consisting of a bundle of n risky assets) with the safe asset. This decision does depend on her subjective risk–return preferences. At a point such as A (Figure I.A.3.9) the *individual* investor is reasonably risk-averse, and holds most of her (dollar) wealth in the safe asset and puts only a little into the market portfolio (in the fixed optimal proportions w_i^*). In contrast Ms B is less risk-averse than Ms A and ends up at B (to the right of M), with a levered portfolio (i.e. she borrows to increase her holdings of the market portfolio in excess of her own initial wealth). An investor who ends up at M is moderately risk-averse, and puts all of her wealth into the market portfolio and neither borrows nor lends at the risk-free rate.

Note that Ms A and Ms B *both* have the same Sharpe ratio as an investor at M – in part this is because all of these investors hold the same risky asset *proportions*. Hence, although Ms B holds more risk $\sigma_B > \sigma_A$ than Ms A, she also expects a higher return $ER_B > ER_A$ – as in Figure I.A.3.9. These two effects just offset each other, so that the expected (excess) return per unit of risk is the same for both Ms A and Ms B and for the investor at M.

I.A.3.9 Summary

It should now be clear how portfolio theory can be used to determine the optimal asset allocation strategy for *any* investor. The portfolio manager first agrees with the investor on the likely future outcomes for expected returns ER_i , variances σ_i and covariances σ_{ij} (or correlations ρ_{ij}). These are likely to be based on historic values plus 'hunches' about particular markets. The analysis might initially be applied only to the *industry asset allocation* problem, namely the optimal proportions to 'place' in the major domestic industries (e.g. chemicals, engineering, services, electrical). The efficient frontier is constructed and, if the investor does not wish to borrow or lend, the analyst can suggest alternative risk–return combinations available along the 'curved' efficient frontier. The investor will choose one of these (e.g. the minimum variance point) based on her own risk–return preferences.

However, if the investor is willing to borrow or lend, she can expand her ‘opportunity set’ by moving up and down the CML. The optimal asset allocation among the risky assets is then given by *her own* ‘market portfolio’ M (and proportions w_i^*). She can then decide whether to put some of her wealth in the risk-free asset or whether to borrow money to provide additional funds to invest in her chosen ‘bundle’ of risky assets – as in the ‘baseline’ case, the latter decision depends on her own level of risk tolerance.

Mean–variance portfolio theory assumes investors can borrow or lend at the risk-free rate. It gives the *optimal proportions* in which risky assets are held. The ‘baseline’ case assumes all investors have the same view about the ‘market-determined variables’, expected returns, variances and covariances. Mean–variance portfolio theory makes the following predictions:

- (i) All investors hold their *risky* assets in the same *proportions* regardless of their preferences for risk versus return. These optimal proportions constitute the market portfolio.
- (ii) Investors’ preferences enter in the second stage of the decision process, namely the choice between the ‘fixed bundle’ of risky securities and the risk-free asset. The more risk-averse the individual, the smaller the proportion of her wealth that will be placed in the bundle of risky assets. A less risk-averse investor will borrow money at the risk-free rate to make additional investments in the risky-assets. All investors hold the same *proportions* in the risky assets but the *dollar amount* in each risky asset depends on the investor’s personal risk tolerance.
- (iii) If investors cannot borrow or lend, they can only choose alternative combinations of risky-asset proportions, along the ‘curved’ efficient frontier. The chosen mix of risky assets *will* depend on the investor’s individual risk tolerance.
- (iv) If investors have different forecasts of expected returns, variances and covariances, they will have their ‘own’ efficient frontier. If they are willing to borrow or lend, then they will choose *their own* particular optimal risky-asset proportions (i.e. their own ‘tangent portfolio’) – these proportions *will differ* for different investors, because the efficient frontier differs for each investor.

Appendix: Mathematics of the Mean–Variance Model

So far we have sidestepped the details of the mathematical calculations needed to obtain the optimal portfolio weights. It turns out that this is not too difficult if there are no constraints put on the optimal weights (e.g. we allow a solution with some $w_i < 0$, i.e. short selling) since we then have a standard quadratic programming problem with an analytic solution. Even in this case, however, some readers might prefer to skip the details in this section, except perhaps for the numerical examples given. If we wish to solve the portfolio allocation problem with constraints (e.g. no short selling) then in general there is no analytic solution and a numerical optimisation routine is needed – these are now commonplace and include add-ons for Excel and programs such as MATLAB, Gauss and Mathematica.

When we allow the investor to borrow or lend at the risk-free rate and also to invest in n risky securities, the optimal solution is the market portfolio (if all investors have the same view of expected returns, variances and covariances). The optimal proportions are determined by the *tangency* of the transformation line with the efficient frontier (short sales are permitted). Mathematically, to obtain the *market portfolio*, we choose the proportions w_i to satisfy

$$\max \theta = \frac{ER_p - r}{\sigma_p} \quad (1)$$

subject to constraints

$$ER_p = \sum_{i=1}^n w_i ER_i \quad (2a)$$

and

$$\sum_{i=1}^n w_i = 1 \quad (2b)$$

where

$$\sigma_p = \left(\sum_{i=1}^n w_i^2 \sigma_i^2 + \sum_{i \neq j} w_i w_j \sigma_{ij} \right)^{\frac{1}{2}} = \left(\sum_{i=1}^n \sum_{j=1}^n w_i w_j \sigma_{ij} \right)^{\frac{1}{2}} \quad (2c)$$

ER_i is the *expected* return on asset i , ER_p and σ_p are the expected return on the portfolio and its standard deviation respectively. The constraint (2a) can be directly incorporated in the maximand (1) by writing θ as:

$$\theta = \frac{\sum_{i=1}^n w_i (ER_i - r)}{\sigma_p} \quad (3)$$

It can be shown (see Chapter II.C) that the ‘first-order conditions’ after differentiating (3) with respect to each w_i in turn can be represented by the following n equations:

$$\begin{aligned} z_1 \sigma_{11} + z_2 \sigma_{12} + \dots + z_n \sigma_{1n} &= ER_1 - r \\ z_1 \sigma_{12} + z_2 \sigma_{22} + \dots + z_n \sigma_{2n} &= ER_2 - r \\ z_1 \sigma_{1n} + z_2 \sigma_{2n} + \dots + z_n \sigma_{nn} &= ER_n - r \end{aligned} \quad (4)$$

where $z_i = \lambda w_i$ and λ is a constant. The constant λ does not affect the solution, since if w_i is a solution to (3) then so is λw_i , since the λ cancels from the numerator and denominator. Having solved equation (4) for z_i then we can determine the optimal values for w_i by noting:

$$\sum_{i=1}^n w_i = 1 = \lambda^{-1} \sum_{i=1}^n z_i \quad (5)$$

hence $\lambda = \frac{\sum_{i=1}^n z_i}{1}$ and therefore $w_i = z_i / \lambda = z_i / \sum_{i=1}^n z_i$.

Since ER_i , r , σ_i^2 and σ_{ij} are known, equation (4) is an n -equation system which can be solved for the n unknowns z_1, \dots, z_n . Equation (4) can be written (for $k = 1, 2, \dots, n$):

$$\sum_{i=1}^n \sigma_{ki} z_i = (ER_k - r),$$

or in matrix notation

$$\mathbf{\Omega} \mathbf{z} = \mathbf{ER} - \mathbf{r} \quad (6)$$

where $\mathbf{z} = (z_1, z_2, \dots, z_n)$, $\mathbf{ER} - \mathbf{r} = (ER_1 - r, ER_2 - r, \dots, ER_n - r)$ and $\mathbf{\Omega}$ is the $(n \times n)$ variance-covariance matrix. It follows that the optimal value for \mathbf{z} is:

$$\mathbf{z}^* = \mathbf{\Omega}^{-1} (\mathbf{ER} - \mathbf{r}) \quad (7)$$

Example: Calculate the optimal weights for the market portfolio for a simple two-variable case where the expected returns and variances of a two-asset portfolio are $ER_1 = 14\%$, $ER_2 = 8\%$, $\sigma_{11} = 36\%$, $\sigma_{22} = 9\%$, $\rho = 0.5$ and the risk-free rate is $r = 5\%$.

Answer: The optimum weights are obtained from the first-order conditions $ER - r = \mathbf{\Omega z}$ or,

equivalently, $ER_i - r = \sum_{i=1}^n \sigma_{ki} z_i$ ($k = 1, 2$ and $n = 2$). We have $\sigma_{12} = \rho \sigma_1 \sigma_2 = 9$ ($= 0.5 \times 6\% \times$

3%) so the first order conditions are:

$$9 = 36z_1 + 9z_2$$

$$3 = 9z_1 + 9z_2$$

The solution is given by $\mathbf{z} = \mathbf{\Omega}^{-1}(ER - \mathbf{r})$, which is:

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{bmatrix} 36 & 9 \\ 9 & 9 \end{bmatrix}^{-1} \begin{pmatrix} 9 \\ 3 \end{pmatrix}$$

Hence (see Chapter II.D) $z_1 = 2/9$ and $z_2 = 1/9$. Thus

$$w_1 = \frac{z_1}{\sum z_i} = \frac{2/9}{1/3} = \frac{2}{3} \text{ and } w_2 = \frac{z_2}{\sum z_i} = \frac{1/9}{1/3} = \frac{1}{3}$$

The expected portfolio returns (using optimal weights) are

$$ER_p = (2/3) 14\% + (1/3) 8\% = 12\%.$$

The standard deviation of the optimal portfolio is

$$\sigma_p = [(2/3)^2(36) + (1/3)^2(9) + 2(2/3)(1/3)9]^{1/2} = \sqrt{21} = 4.58\%.$$

The optimal portfolio weights are $w_1 = 2/3$ and $w_2 = 1/3$, and therefore the optimum ‘market portfolio’ has an expected return $ER_p = 12\%$ and a standard deviation of $\sigma_p = 4.58\%$, which gives the highest Sharpe ratio possible. The optimal portfolio is the so-called ‘tangent portfolio’. That is, it provides optimal weights to hold in the risky assets so that you maximise the Sharpe (reward to variability) ratio.

I.A.4 The CAPM and Multifactor Models

Keith Cuthbertson and Dirk Nitzsche¹

The capital asset pricing model (CAPM) is central in determining the average return required by investors for holding a particular stock (or portfolio of stocks). It therefore determines the cost of equity capital for a firm raising funds when issuing more shares, and this ‘cost’ is used to discount the future cash flows expected from a risky physical investment project, in order to decide whether the project is viable. Hence the CAPM is useful for issues in corporate finance. What use is the CAPM for an investor? In the CAPM, the riskiness of a stock when held as part of a well-diversified portfolio is determined by how much the stock adds to overall *portfolio* risk, and this is measured by the stock’s ‘beta’. When the market as a whole goes up or down by 10% and the return on a particular stock then moves by plus or minus 20%, then the beta of the stock equals 2. A stock with a beta of 0.5 is therefore less risky than one with a beta of 2. You can reduce the overall riskiness of a portfolio of stocks by selling ‘high beta’ stocks and using the funds to purchase ‘low beta’ stocks. However, you might incur high transaction costs (e.g. bid–ask spreads, brokerage fees) especially if you just wanted to change your ‘risk exposure’ for a relatively short period (e.g. 3–6 months). Once you know the beta of your stock portfolio, we see in later chapters how this ‘risk reduction’ can be achieved more easily and cheaply by using stock index futures. Every financial institution in developed economies has to hold capital against the riskiness of its assets (and liabilities) and, as we see in later chapters, the beta of a stock can be used in calculating the financial intermediaries’ ‘dollar’ exposure to risk (called its value at risk) on its stock portfolio (in domestic and foreign assets). The CAPM is therefore a key ‘tool’ in analysing and solving a wide range of important practical problems.

In this chapter we analyse the (basic one-period) CAPM, the single-index model and multifactor models such as the arbitrage pricing theory (APT). The CAPM is widely used in the finance literature to determine the average return required by shareholders on a particular asset based on its contribution to overall portfolio risk. We also present a brief account of the APT, which relates the expected return on a security to a *set* of variables called ‘factors’, which could include market-wide effects due to interest rates, exchange rates, etc. Throughout this chapter we consider that the only risky securities are equities (stocks), although strictly the model applies to choices among *all* risky assets (e.g. stocks, bonds, real estate).

¹Keith Cuthbertson is Professor of Finance and Dirk Nitzsche is Senior Lecturer at the Cass Business School, City University, London.

I.A.4.1 Overview

The CAPM provides an elegant model of the determinants of the equilibrium expected or required return on any *individual* risky asset. It predicts that the expected return on a risky asset ER_i consists of the risk-free rate r plus a risk premium (rp_i):

$$ER_i = r + rp_i = r + \beta_i(ER_m - r) \quad (\text{I.A.4.1})$$

where the risk premium, $rp_i = \beta_i(ER_m - r)$ and $\beta_i = \text{cov}(R_i, R_m)/\text{var}(R_m)$. The risk premium is proportional to the excess market return ($ER_m - r$) with the constant of proportionality given by the beta (β_i) of the individual risky asset. The excess market return ($ER_m - r$) is also known as the *market risk premium* since it is the additional average return on the market portfolio over and above the risk-free rate. It is a ‘payment’ for holding the risky market portfolio. The definition of security i ’s beta, β_i , indicates that it:

- depends positively on σ_{im} , the covariance between the return on security i and the market portfolio, $\text{cov}(R_i, R_m)$;
- is inversely related to the variance σ_m^2 of the market portfolio, $\text{var}(R_m)$.

Loosely speaking, if *ex post* (or actual average returns) approximate the *ex ante* expected return ER_i , then we can think of the CAPM as explaining the average monthly return (over, say, a 36-month period) on security i .

What does the CAPM tell us about equilibrium-required returns on individual securities in the stock market? First, note that $(ER_m - r) > 0$, otherwise no risk-averse investor would hold the market portfolio of risky assets when she could earn more, *for certain*, by investing all her wealth in the risk-free asset. The CAPM predicts that for those stocks that have a zero covariance with the market portfolio, they will be willingly held as long as they have an expected return equal to the risk-free rate (put $\beta_i = 0$ in equation (I.A.4.1)).

Second, returns on individual stocks tend to move in the same direction and hence, in general, $\text{cov}(R_i, R_m) \geq 0$ and $\beta_i \geq 0$. Securities that have a large positive covariance with the market return ($\beta_i > 0$) will have to earn a relatively high average return. As we have seen in the previous chapter, this is because the addition of such a security to the portfolio does little to reduce *overall portfolio* variance, and to offset the latter you require a high average return on this security.

The CAPM also allows one to assess the relative volatility of the *expected* returns on individual stocks on the basis of their β_i values. Stocks for which $\beta_i = 1$ have a return that is expected to move one-for-one with the market portfolio (i.e. $ER_i = ER_m$) and are termed ‘neutral stocks’. If $\beta_i > 1$ the stock is said to be an ‘aggressive stock’, since, on average, it moves *more* than changes in

the expected market return (either up or down). Conversely, ‘defensive stocks’ have $\beta_i < 1$. Therefore investors can use betas to rank the relative ‘safety’ of various securities and can combine different shares to give a desired beta for the portfolio.

I.A.4.2 Capital Asset Pricing Model

The CAPM is a logical consequence of mean–variance portfolio theory and assumes:

- all investors have homogeneous expectations;
- investors choose their risky asset proportions by maximising the Sharpe ratio (see I.A.3);
- investors can borrow or lend unlimited amounts at the risk-free rate;
- the market is in equilibrium at all times.

It provides an equation to determine the return required by investors to willingly hold any particular risky asset (as part of a well-diversified portfolio):

$$\text{Required return on asset } i = \text{Risk-free rate} + \text{Risk premium}$$

or, in symbols,

$$ER_i = r + \beta_i(ER_m - r) \tag{I.A.4.2}$$

where β_i is the asset’s beta. The ‘risk premium’, therefore, consists of the market risk premium ($ER_m - r$), which is the same for all securities, multiplied by ‘beta’, which is sometimes referred to as the *market price of risk* (more accurately, the ‘price of market risk’). The latter terminology will appear in the APT and for the moment we will not use it for the CAPM, since it does not add a great deal at this point.

Let us see how the CAPM arises from portfolio theory. In order that the efficient frontier be the same for *all* investors, they must have homogeneous expectations about the underlying market variables ER_i , σ_i^2 and σ_{ij} . With homogeneous expectations, *all* investors hold *all* the risky assets in the proportions given by the tangency point M , the market portfolio. Because all n assets are held at M , there is also a set of equilibrium expected returns ER_i (for the n assets) corresponding to point M . The CAPM equation representing the equilibrium returns for asset i takes account of the fact that, when held as part of a portfolio, asset i might reduce the risk of the overall portfolio (i.e. low or negative covariances again), hence:

The riskiness of asset i , when *considered as part of a diversified portfolio*, is not its own variance σ_i^2 but the covariance between R_i and the market return R_m .

To put the CAPM intuitively, ask yourself the following question. Would you be willing to hold *only* the single risky security A, which has a return standard deviation of $\sigma_A = 60\%$ p.a. but a rather low average return of only 5.8% p.a.? If this is the *only* security you hold then you would most likely want a higher average return, given its high *individual* risk of 60% p.a. Here, this single security constitutes your total portfolio. Now, suppose you *already hold* a number of risky assets and you are considering adding A to your existing portfolio. What is more, assume that A's return has a rather low covariance with the asset returns in your existing portfolio. The latter implies that if you include A in your portfolio the overall increase in *portfolio* risk will be very small. Hence, the *incremental* risk of adding A to your existing portfolio is small and therefore you might be willing to hold A, even though it has a relatively low average return of only 5.8%. If you hold a diversified portfolio, it is the incremental contribution of A to the *overall* portfolio risk that is important in determining the average return for A and not the risk of this asset considered in isolation. Indeed, in Chapter I.A.3 we noted that A's specific risk could be costlessly diversified away. Finally, suppose, for example, that A has a beta of 0.1, the average excess return on the market is 8% p.a. and the risk-free rate $r = 5\%$ p.a. Then the required average return, according to the CAPM, would be exactly 5.8% p.a. – so 'A' does earn a return that just compensates for its incremental contribution to overall portfolio risk, even though its 'own risk' is very high (i.e. $\sigma_A = 60\%$).

The reason why stock A earns only 5.8% even though it has a high standard deviation of 60% is that much of the individual risk of this stock can be diversified away almost costlessly, by including this stock along with other stocks, in a diversified portfolio. Because most of the variability of 60% in the return of stock A is removed by including stock A in your existing portfolio, you should receive no reward (i.e. average return) based on stock A's high individual standard deviation. The CAPM implies that you should only receive a reward, i.e. an average return, based on how much stock A contributes to the risk of your *whole* portfolio. The contribution of stock A to your overall portfolio risk, is how much this stock moves relative to changes in the overall market return. If stock A has a beta greater than 1 it moves much more than the market return (either up or down) and the CAPM predicts that this stock will be willingly held by investors only if it has a high average return.

I.A.4.2.1 Estimating Beta

An estimate of an asset's beta can be obtained using an ordinary least-squares *time series* regression (see Chapter II.F):

$$(R_i - r)_t = \alpha_i + \beta_i (R_m - r)_t + \varepsilon_{it} \quad (\text{I.A.4.3})$$

where $\beta_i = \text{cov}(R_i, R_m) / \sigma_m^2$ is the formula for the ordinary least-squares estimate of beta. For example, if the monthly return on asset i , in excess of the risk-free rate $(R_i - r)_t$ over, say, the last 60 months, is regressed on the monthly excess return on the market $(R_m - r)_t$, then the slope gives an estimate of β_i . In practice some aggregate stock index such as the S&P 500 or the FT Actuaries Index is used as a measure of the return on the market portfolio, R_m . If the CAPM is the correct model of equilibrium returns then we expect the estimate of the intercept α_i to be ‘close to’ zero.

I.A.4.2.2 Beta and Systematic Risk

The CAPM predicts that only the covariance of returns between asset i and the market portfolio influences the average excess return on asset i . No additional variables such as the dividend price ratio, the size of the firm or the price–earnings ratio should influence average excess returns on a stock. All contributions to the risk of asset i , *when it is held as part of a diversified portfolio*, are summed up in its beta. The beta of a security represents that part of risk that cannot be diversified away, hence: *Beta represents an asset’s systematic (market or non-diversifiable) risk and is the only source of risk that contributes to the excess return of asset i .*

Of course, an individual investor may choose not to use the full mean–variance optimisation procedure in determining her portfolio allocation. However, ‘beta’ may still be useful. For example, if an investor requires a portfolio with a specific beta (e.g. ‘aggressive portfolio’ with $\beta_p > 1$) then this can be easily constructed by combining securities with different betas in different proportions, where:

$$\beta_p = \sum w_i \beta_i \tag{I.A.4.4}$$

Of course, the weights w_i here do not represent the *optimal* weights given by mean–variance portfolio theory but simply those proportions held by this particular investor.

I.A.4.3 Security Market Line

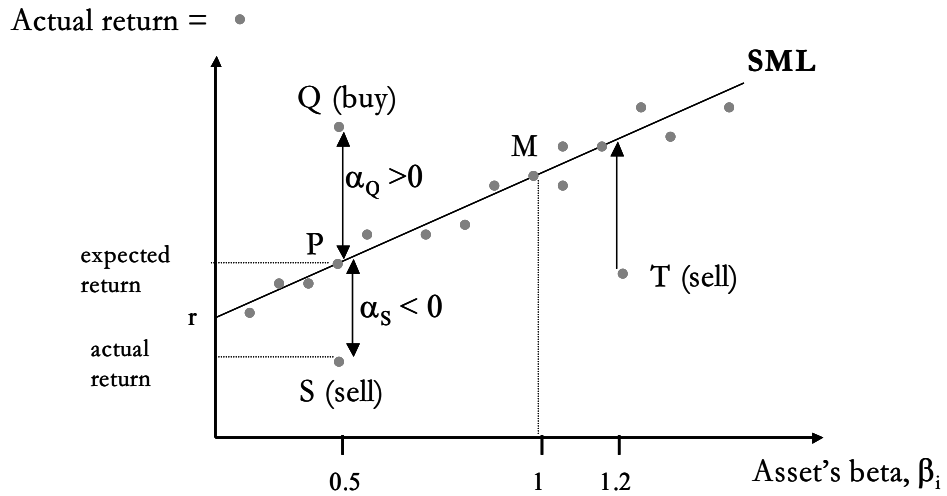
The CAPM can be ‘rearranged’ and expressed in terms of the security market line (SML). Suppose that the historic average value of the market risk premium $(R_m - r)$ is 8% p.a. and the risk-free rate is $r = 5\%$, then the CAPM becomes:

$$\bar{R}_i = 5 + 8\beta_i \tag{I.A.4.5}$$

This linear relationship between the average return \bar{R}_i and β_i is known as the security market line:

Figure I.A.4.1

Required return = SML



Securities which lie above (below) the SML have a positive (negative) 'alpha' indicating an 'abnormal return' after correcting for 'beta risk'.

The SML (Figure I.A.4.1) relates \bar{R}_i on security i with its beta β_i , and if the CAPM is correct then all securities should lie on the SML. According to the CAPM/SML, the average monthly excess return (say, over the last 5 years) on each asset ($\bar{R}_i - r$) should be proportional to that asset's beta β_i . The SML is therefore a *cross-section (regression) relationship across different stocks* (whereas each beta is estimated using a time series of data).

Given the definition of β_i in equation (I.A.4.1), we see that the market portfolio M has a beta of 1, since $\text{cov}(R_i, R_m) = \text{var}(R_m)$ if asset i is the market portfolio. If a security has $\beta_i = 0$ then it will earn an average return equal to the risk-free rate. A security with a high beta has a high covariance (correlation) with the market return and therefore adds considerable incremental risk to the portfolio as a whole. Hence, such a security should earn a high average return to compensate for this incremental portfolio risk. Mathematically, the average return on two securities is proportional to their relative betas— the average excess return on security i divided by that on security j :

$$\frac{\bar{R}_i - r}{\bar{R}_j - r} = \frac{\beta_i}{\beta_j}$$

The SML can be used to try to pick underpriced and overpriced stocks. To see this, consider a security S (Figure I.A.4.1) with a beta of 0.5. You could duplicate the beta of security S by buying a portfolio with 50% in the safe asset (with $\beta = 0$) and 50% in a single security with $\beta = 1$ (i.e. β_P

$= 0.5 \times 0 + 0.5 \times 1 = 0.5$). But this ‘synthetic portfolio’ would lie on the SML (at P), has a higher expected return and hence would dominate S . Hence S would be sold, since its actual return is less than its equilibrium return given by the SML. If S were sold, its *current* price would fall and this would raise its *expected* return, so that S moves towards P . (Similarly, a security T with $\beta = 1.2$ could be duplicated by borrowing 20% of your wealth at the safe rate and using your own funds plus borrowed funds to invest in a security with a $\beta = 1$.)

Alternatively, consider a security like Q also with $\beta_i = 0.5$ (Figure I.A.4.1), but which currently has a higher average return than indicated by the SML. An investor should purchase Q . Securities like Q and S are currently mispriced (i.e. they are not on the SML) and a speculator might short-sell S and use the funds to purchase Q . If the mispricing is corrected then the price of Q will rise as everyone seeks to purchase it, because of its current high ‘abnormal’ average return. Conversely, everyone seeks to short-sell S , so its price falls in the market. If you spotted this mispricing first and executed your trades before everyone else, then you would earn a handsome profit from this mispricing. A subtle point is that, if the market return ($R_m - r$) either rises or falls unexpectedly over this period, you still make a profit. (Can you see why? *Hint*: both S and Q have the same beta.) To implement this active long-short strategy one has ‘to graph’ the average historic return for a set of securities \bar{R}_i (say monthly returns averaged over the past 5 years) against their β_i estimates and look for ‘big outlier securities’ like Q and S . However, remember that in practice this investment strategy is risky since it assumes the CAPM is ‘true’, β_i is measured correctly and that any mispricing will be corrected over a reasonably short time horizon.²

I.A.4.4 Performance Measures

Suppose the performance of fund managers is assessed every 3 months. Over a five-year horizon two fund managers might have identical average (excess) returns of 10% p.a. from their ‘active’ portfolio strategies. If fund manager A has three-monthly returns (expressed at an annual rate) of 10.1%, 9.8%, 10.2%, 9.9% etc., whereas fund manager B has a sequence of returns like 25% , –15%, 40% , –10%, which fund manager would you be happier with? If you have a three-month horizon and you like return but dislike risk, you would prefer B. This simple example demonstrates that most investors are concerned not only with average return but also with the risk associated with a particular investment strategy. There is no unambiguous measure of risk. In

² There are many studies that examine the validity of the CAPM. Although not perfect, the CAPM provides a useful baseline model to explain the return required by investors on particular stocks and portfolios of stocks. The CAPM can be extended to include ‘other factors’ as well as the market return, and such multifactor models explain the average returns on a wide range of alternative portfolios. For a survey of tests of the CAPM and multifactor models see Cuthbertson and Nitzsche (2004), and for a more formal mathematical treatment see Cochrane (2001).

the above example it is clear that the standard deviation of returns is larger for manager B than for manager A, and the Sharpe ratio measures the ‘return per unit of risk’.

What is causing the differential movement in the returns of managers A and B? It may be that B has a portfolio of stocks with a very high beta whereas A’s portfolio has a rather low beta. Since both managers earned the same average return we might conjecture that B is earning less return per unit of ‘beta risk’ than A. Here we are using the beta of the portfolio as a measure of the (relative) riskiness of the two fund managers – this is the basis of Jensen’s ‘alpha’ as a measure of risk-adjusted performance. What is important is to measure the performance of fund managers (or other investment professionals) after making some allowance for the riskiness of their investment strategy. The CAPM provides the basis for one approach, which is widely used in assessing investment performance. (The approach can be extended to multifactor models but we do not pursue this here.)

The CAPM/SML predicts that the excess return on any stock adjusted for the exposure to ‘market risk’ on that stock β_i should be the same for all stocks (and all portfolios of stocks). Algebraically this may be expressed as:

$$(ER_i - r)/\beta_i = (ER_j - r)/\beta_j = \dots = ER_m - r \quad (\text{I.A.4.6})$$

Therefore, the CAPM/SML predicts that the average return on *all* stocks (and portfolios of stocks) corrected for ‘beta risk’ should be the same and that the intercept in the time series regression (I.A.4.3) should be zero. Of course, equation (I.A.4.6) applies under the somewhat restrictive assumption of the standard CAPM that the market is in equilibrium at all times. In the real world, however, it is possible that over short periods the market is not in equilibrium and profitable opportunities arise. This is more likely to be the case if investors have divergent expectations, or if they take time to learn about a new environment that affects the returns on stocks of a particular company, or if there are some ‘irrational’ agents who base their investment decisions on what they perceive are ‘trends’ in the market and do not correctly account for risk.

I.A.4.4.1 Sharpe Ratio

Any useful performance index has to consider the return *relative to* the risk of the portfolio and then rank alternative portfolios accordingly. We have already met (see Chapter I.A.3) Sharpe’s *reward-to-variability ratio*, which is defined for portfolio *i* as:

$$S_i = \frac{ER_i - r}{\sigma_i} \quad (\text{I.A.4.7})$$

where ER_i is the expected (average) return on portfolio i , σ_i is the standard deviation of portfolio i , and r is the risk-free rate. Here risk is measured by the standard deviation of portfolio returns. Usually the Sharpe ratio is used to measure performance where you already hold cash and are considering investing in two (or more) alternative portfolios, X and Y . You observe the historic returns (e.g. over the last 60 months) on the portfolios and compute the average excess returns $(\bar{R}_i - \bar{r})$ and their standard deviations σ_X and σ_Y . The portfolio with the highest (historic) Sharpe ratio has performed the ‘best’. The Sharpe ratio of a passive investment strategy such as holding a tracking portfolio on the S&P 500 is also usually used as a benchmark against which to judge the alternative ‘active’ (or stock-picking) strategies X and Y .

I.A.4.4.2 Jensen’s ‘alpha’

Jensen’s performance index is based on the CAPM and is the intercept α_i in the following regression:

$$(R_i - r)_t = \alpha_i + \beta_i(R_m - r)_t + \varepsilon_{it} \quad (\text{I.A.4.8})$$

To run this regression we need time series data on the excess return on the market portfolio and the excess return on the chosen portfolio i . For example, Jensen’s alpha could be used to assess the stock-picking skill of a fund manager. The dependent variable would then be the excess returns of the relevant fund. We then obtain estimates for α_i and β_i , where:

$$\hat{\alpha}_i = (\bar{R}_i - \bar{r}) - \hat{\beta}_i(\bar{R}_m - \bar{r}) \quad (\text{I.A.4.9})$$

It is immediately apparent from equation (I.A.4.8) that if $\alpha_i = 0$ then we have the standard CAPM/SML and this security/portfolio would lie on the SML. But if $\alpha_i > 0$ the fund i earns a return in excess of that given by the CAPM/SML and would be in a position like Q (Figure I.A.4.1), where $\alpha_Q > 0$. For $\alpha_i < 0$ the fund manager has underperformed relative to the required rate of return given by the SML (see point S in Figure I.A.4.1). The ‘alpha’ is a measure of ‘how far’ the average return is from the SML and indicates abnormal returns (in the past) for this particular asset (or portfolio of assets). Hence Jensen’s alpha actually measures the *abnormal return* on the asset after correcting for the *incremental* or ‘beta risk’ of the assets.

Jensen’s alpha for two alternative portfolios X and Y is obtained from a time series regression of $(R_i - r)_t$ on $(R_m - r)_t$ where $i = X$ or Y . The ‘best’ fund is that which has the highest average abnormal return as measured by its alpha. Jensen’s alpha is easily generalised to cases where there are more sources of risk than the market return (see Section I.A.4.6 below).

Both the Sharpe ratio and Jensen’s alpha are used to rank alternative investment strategies, and often (but not always) they give similar rankings.³ Their widespread use is due to the fact that they both measure returns corrected for risk and are easy to implement.

I.A.4.5 The Single-Index Model

The single-index model (SIM) is not really a ‘model’ in the sense that it embodies any behavioural hypotheses (e.g. about the return required to compensate for holding ‘market risk’) but it is merely a *statistical assumption* that the return on *any* security R_{it} may be adequately represented as a linear function of a single (economic) variable I_t (e.g. inflation or interest rates):

$$R_{it} = \theta_i + \delta_i I_t + \varepsilon_{it} \quad (\text{I.A.4.10})$$

where ε_{it} is a random error term and equation (I.A.4.10) holds for any security (or portfolio) $i = 1, 2, \dots, n$ and for all time periods. Hence I_t could be any variable that is found to be correlated with R_{it} , and the SIM has no specific theoretical model that seeks to explain this observed correlation.

Clearly, even if the variable chosen for I_t is the excess market return $R_m - r$ the equations for the SIM and the CAPM (I.A.4.1) do differ, because the CAPM uses *excess* returns on asset i and has a zero intercept term. However, in some of the literature the CAPM and SIM are often treated as equivalent (which they are not) and the CAPM is referred to as a ‘single-factor model’. The latter is acceptable provided I_t is the *excess* market return $R_m - r$ and the dependent variable is the *excess* return on the stock, $R_i - r$. In any case we have deliberately denoted the coefficient on I_t as δ_i rather than β_i to emphasise the fact that in general the SIM differs from the CAPM.

Equation (I.A.4.10) can be estimated by ordinary least-squares regression (or more sophisticated techniques) using, say, 60 months of time series data. This gives estimates of the parameters and other ‘diagnostic output’ such as the R-squared of the regression.

Strictly speaking, the SIM also assumes that ε_{it} , the ‘unexplained’ or ‘residual’ or ‘idiosyncratic’ element of the return for any security i , is independent of I_t and of the residual return for any other security j . Then it is easy to show that:

³ There is no unambiguously superior measure of risk. Jensen’s alpha ‘adjusts’ the observed average portfolio return for the ‘beta risk’ of the portfolio and therefore assumes your chosen portfolio is a subset of the market portfolio. The Sharpe ratio measures risk by the standard deviation of your portfolio and therefore assumes your portfolio contains all the assets in the market portfolio (but held in different proportions to the optimal market portfolio weights, as your active strategy involves ‘stock picking’). For a discussion of the practical issues when using these two measures see Cuthbertson and Nitzsche (2001, 2004), and for a more technical discussion see Cerny (2004).

$$\bar{R}_i = \theta_i + \delta_i \bar{I} \quad (\text{I.A.4.11})$$

$$\sigma_i^2 = \delta_i^2 \sigma_I^2 + \sigma_{\epsilon_i}^2 \quad (\text{I.A.4.12})$$

$$\sigma_{ij} = \delta_i \delta_j \sigma_I^2 \quad (\text{I.A.4.13})$$

Equation (I.A.4.11) simply says that the average return on stock i depends on the average value of the index \bar{I} and the value of δ_i and θ_i for that stock. Shares that have more exposure to the index (i.e. a higher $\delta_i > 0$), will have a higher average return (for $\bar{I} > 0$). From (I.A.4.12) we see that, for any individual security, the SIM can be used to apportion the *total* volatility of its return σ_i^2 into that due to

- its *systematic risk*, caused by changes in the index I_t (which is common across all securities); and
- its *specific risk* $\sigma_{\epsilon_i}^2$, caused by random events that affect only this particular security.

In other words, (I.A.4.12) implies, for any security i :

$$\text{Total variance of return} = \text{'delta'} \times \text{variance of index} + \text{variance of specific risk}.$$

In general, for monthly data for a particular stock the R^2 of (I.A.4.10) will be rather low. This would imply that the variability of the individual stock returns is *not* particularly well explained by movements in the index. Suppose figures for σ_i , $\delta_i \sigma_I$, and σ_{ϵ_i} are 40%, $0.9 \times 15\% = 13.5\%$ and 26.5%, respectively (with the volatilities being expressed in annual terms, as usual). Then much of the total monthly return variability (here 40%) for individual stocks would be due to specific risk (26.5%), rather than market risk (13.5%). Of course portfolio theory points out that this specific risk can be diversified away at near zero cost if this security is held as part of a well-diversified portfolio. That is why the CAPM/SML predicts that the average return on any individual stock depends not on its 'own variance' but on its covariance (correlation) with the rest of the stocks in the portfolio.

Equation (I.A.4.13) shows that in the SIM the covariance between any two securities depends only on the values of δ_i , δ_j and σ_I . This is useful when calculating the value at risk of a portfolio of stocks (see Chapter III.A.2).

The independence assumption across different securities rarely holds. Particularly when the stocks are within the same country index, it is unlikely that 'shocks' or 'news' that influence returns on firm A will not also sometimes influence the returns on firm B. When comparing returns in different countries the SIM has somewhat greater applicability, since macroeconomic shocks (e.g. unexpected changes in interest rates) may not be synchronised across countries.

However, unbiased estimates of the parameter can be obtained even when the residual returns are correlated and therefore the SIM is quite widely used in practice. Also, it can be ‘improved’ by extending it to a ‘multi-index’ model by including more variables that are thought to influence all stock returns (to a greater or lesser extent) – for example, macroeconomic variables such as interest rates or exchange rates or ‘factor-mimicking’ variables such as the returns on ‘high minus low book-to-market value’ shares.⁴ Such multifactor models could be used to pick undervalued and overvalued stocks, in the same way as we used the SML previously.

I.A.4.6 Multifactor Models and the APT

For most people the arbitrage pricing theory is a little difficult to grasp as its derivation is based on some theoretical ideas on the impact of portfolio diversification on the average return on any stock, which are not as intuitively appealing as the CAPM. Unlike the CAPM, the APT allows a number of potential variables (factors) to influence the expected return on any asset i . The CAPM has only one factor, namely the excess market return. Broadly speaking, the APT allows the return R_{it} on asset i to be influenced by a number of market-wide variables or ‘factors’, such as interest rates and the exchange rate. The sensitivities of the return on asset i to each of these factors are known as the ‘factor betas’. The APT leads to a regression model:

$$R_{it} = a_i + \sum_{j=1}^k b_{ij} F_{jt} + \varepsilon_{it} \quad (\text{I.A.4.14})$$

where F_j is the j th factor (variable), b_{ij} is the beta of the j th factor and ε_{it} is a random error. These factor betas are of course conceptually different from ‘the beta’ in the CAPM. Using a relatively sophisticated proof based on risk-free arbitrage, it is possible to show that equation (I.A.4.15) gives an explicit expression for the *equilibrium return* ER_i on any risky asset:

$$ER_i = \lambda_0 + \lambda_1 b_{i1} + \lambda_2 b_{i2} + \dots + \lambda_k b_{ik} \quad (\text{I.A.4.15})$$

Equation (I.A.4.14) is similar to the CAPM equation if we assume there is only a single factor that is the excess return on the market portfolio. Similarly, equation (I.A.4.15) is the APT

⁴ For example, a factor-mimicking portfolio might be a portfolio that contains those stocks (from the universe of stocks on the NYSE) that have the highest book-to-market ratio (e.g. the highest decile). These are known as ‘value stocks’. As the book value of the firm does not vary very much compared with the market value (i.e. price \times number of stocks), high book-to-market implies stocks with low prices. When we examine these stocks we find that the reason why they have low prices is that they are in some form of financial distress (e.g. high interest payments relative to profits). This ‘distress factor’ is common to all these stocks and it makes them relatively risky (i.e. they are close to bankruptcy) and hence we might expect that the return required by investors to hold such stocks would be relatively high. In this sense stocks with a high book-to-market ratio ‘mimic’ the high default risk of such stocks. See Cuthbertson and Nitzsche (2004) for an analysis of value-growth, momentum and other stock-picking ‘styles’.

equivalent of the SML, since it shows that the expected return on any asset i depends linearly on a set of (factor) betas.

Note that in the APT a particular factor can influence *actual* returns but that does not necessarily imply it has an effect on *equilibrium* returns. That is, its λ_j value may be zero (see below). If a factor does not influence equilibrium returns, the factor is said to be ‘not priced’. The λ_j is sometimes referred to as ‘the price of beta risk’ for the factor F_j .

Assume we have three factors – ‘inflation’, ‘interest rates’ and ‘economic output’ – with only two factors influencing equilibrium returns (i.e. only inflation and interest rates are priced):

$$R_{it} = a_1 + b_{i1}F_{1t} + b_{i2}F_{2t} + b_{i3}F_{3t} + \varepsilon_{it} \quad (\text{I.A.4.16})$$

$$ER_i = \lambda_0 + \lambda_1 b_{i1} + \lambda_2 b_{i2} \quad (\text{I.A.4.17})$$

where R_{it} is the actual return on stock i	ER_i is the equilibrium return on stock i
b_{ij} is the ‘risk’ factor weights	F_{1t} is the change in inflation
F_{2t} is the change in interest rates	F_{3t} is the change in output

The factors F_{it} are measured as changes rather than level terms, so strictly speaking they are unexpected events or ‘surprises’. In practice, the *change* in a variable is often deemed to be unforecastable and is taken as a measure of an unexpected event. Notice that the APT implies that the λ s are the same for all stocks (but the b_{ij} s differ across stocks). The theory therefore assumes that stocks with the same sensitivity to the economic factors (i.e. the same b_{ij} s) will offer the same equilibrium return. This is because the factors that are ‘priced’ give rise to ‘risk’ that cannot be diversified away and therefore must influence the risk-adjusted (required) return on the stock.

I.A.4.6.1 Portfolio Returns

Specific risk can be diversified away by holding a large number of securities. For simplicity, suppose there is only one systematic risk factor, F , n securities in the portfolio and these are held in proportions w_i ($i = 1, 2, \dots, n$). Then

$$R_p = \sum_{i=1}^n w_i R_i + \sum_{i=1}^n w_i (ER_i + b_i F + \varepsilon_i) = \sum_{i=1}^n w_i ER_i + \left(\sum_{i=1}^n w_i b_i \right) F + \sum_{i=1}^n w_i \varepsilon_i \quad (\text{I.A.4.18})$$

Thus, the return on the portfolio is a weighted average of the expected return *plus* the weighted average of the beta for each security (multiplied by the factor) *plus* a weighted average of the specific returns. If specific returns are uncorrelated across securities, some will be positive and

some negative, but their weighted sum is likely to be close to zero. In fact, as the number of securities increases the last term on the right-hand side of equation (I.A.4.18) will approach zero and the specific risk will have been diversified away. Hence, in general, the APT predicts that:

The ‘actual’ return on a portfolio is made up of the expected returns on the individual securities and the systematic risk as represented by the economy-wide news ‘factors’.

So far we have an equation for actual portfolio returns. In a rather complex no-arbitrage proof, Ross (1976) shows that the *equilibrium expected return* on any asset i depends linearly on the factor betas, as in equation (I.A.4.15). Hence, in principle, the APT is more general than the CAPM since it allows several ‘factors’ to influence asset returns.

Such multifactor models can be estimated using monthly excess returns once one has isolated the important factors. The intercept in these models, Jensen’s alpha, is often used as a measure of abnormal performance where we allow several factors to influence the returns on different securities or portfolios.

I.A.4.7 Summary

The CAPM implies that, in equilibrium, the expected excess return on any risky asset is proportional to the excess return on the market portfolio. The constant of proportionality is called the asset’s ‘beta’. Beta provides a measure of the risk of an individual security when it is held as part of a diversified portfolio. The higher the correlation between the stock return and the market return, the more the stock contributes to overall portfolio risk and hence the higher its average return.

The single-index model assumes that any risky asset return depends on a single ‘index’ or ‘factor’. Estimation of the SIM provides a measure of the relative importance of the asset’s systematic risk to its specific risk. The APT is a multifactor model of equilibrium asset returns based on arbitrage. There are two key features of the APT:

- *Actual* returns may depend on a number of market-wide variables or ‘factors’ (e.g. interest rates and inflation), each of which has its own ‘beta’ coefficient.
- *Equilibrium* returns on asset i depend on a weighted average of its factor ‘betas’. These weights λ are the same for all assets and are termed ‘the price’ of each risk factor.

References

- Cerny, A (2004) *Mathematical Techniques in Finance* (Princeton, NJ: Princeton University Press).
- Cochrane, J H (2001) *Asset Pricing* (Princeton, NJ: Princeton University Press).
- Cuthbertson, K, and Nitzsche, D (2001) *Investments: Spot and Derivative Markets* (Chichester: Wiley).
- Cuthbertson, K, and Nitzsche, D (2004) *Quantitative Financial Economics: Stocks, Bonds and Foreign Exchange* (Chichester: Wiley).
- Ross, S.A. (1976) 'The Arbitrage Theory of Asset Pricing', *Journal of Economic Theory*, Vol. 13, pp. 341–360.

I.A.5 Basics of Capital Structure

Steven Bishop¹

I.A.5.1 Introduction

How should a business be funded? How much debt? What form of debt? How much equity? Does it matter? Answering these questions is at the heart of capital structure choice. The answer must be assessed relative to the overall objective of a business enterprise, and we view this as being to maximise the value of the business for its shareholders. Consequently, capital structure choice requires an understanding of the relationship between capital structure and business value.

This chapter focuses on the capital structure decision primarily from the perspective of an industrial company determining the most appropriate financing for its long-term assets and working capital. This is relevant for a lending institution considering the appropriate level of debt for its customers.

Financial institutions, particularly banks, face the same issues when considering their own capital structure. There are, however, additional complexities. Banks have a unique perspective on funding because of their role as deposit-taking institutions. Deposits are a source of low-cost funding, but are, at the same time, a 'product' that is offered to the public.

In addition, the indirect costs of financial distress for a bank are likely to be significantly larger than for an industrial company and, as a result, there are regulatory requirements that influence the capital structure decision. That is, there are minimum equity requirements and strict restrictions on the extent to which hybrid securities may qualify as capital for regulatory purposes. For most banks, however, these regulatory constraints are not binding; banks choose to hold more capital than regulators require. It is likely that while the specifics may vary, the general issues affecting capital structure for banks mirror those of other listed companies. The specific issues relevant for determining risk capital for banks are discussed in Section III of *The PRM Handbook*.

A key determinant of capital structure choice for any business is the risk or variability of the operating earnings stream. Generally, the higher this risk the less debt can be supported, essentially due to the exposure to distress and bankruptcy costs. Many firms will select a desired credit rating or probability of distress based on the volatility of the earnings stream and use this

¹ Visiting Fellow, Macquarie Applied Finance Centre, Macquarie University; Director and Principal, Education and Management Consulting Services Pty Ltd.

to establish the interest bill, or amount of debt that can be supported. Further, hedging risks can reduce risk and enable additional debt financing and adding value. We expect banks to have lower operating earnings volatility than industrials, and therefore higher debt levels.

We know that funding decisions affect the behaviour of managers under certain circumstances and that this will affect the value of a business. Exactly what the best funding mix is for a particular business is a little imprecise. There are, however, a number of key guidelines that we can provide to help directors make decisions in this area that create shareholder value. The objectives of this chapter are to identify differences between debt and equity and present the drivers of choice between them for a corporation.

As noted above, the underlying assumption in the analysis is that the primary objective of a business is to maximise its value for shareholders. For the purposes of this chapter we will view maximising the value of the operating cash-flow stream as also maximising the value of the equity cash-flow stream.²

The value we focus on is the present value of operating free cash flows over time.

$$V = \sum_{t=1}^{\infty} \frac{OFCF_t}{(1 + WACC)^t} \quad (\text{I.A.5.1})$$

where

- V is the value of the business and is equal to the value of debt (D) plus the value of equity (E);
- $OFCF_t$ is expected operating free cash flow after tax in period t (i.e., revenue less operating expenses and capital expenditure), that is, it is the mean of a probability distribution of possible operating free cash-flow outcomes in period t ; and
- $WACC$ is the weighted average cost of capital and is equal to the weighted average of the cost of debt and the cost of equity,

² The value of the firm (V) can be viewed as the sum of the value of debt (D) plus the value of equity (E), that is, $V = D + E$. Since there is a limit to the claim debt has on the value of the company, maximising V is effectively maximising E . However, E could be maximised by transferring value from D . Thus maximising V and E might not be the same. In fact the text will examine ways management might transfer wealth from D to E and, in so doing, provide explanations for the existence of covenants, aspects of corporation law and the reason for auditing of the books. Nevertheless, our primary focus is on how V changes with changes in the debt–equity mix because of the contractual arrangements which minimise wealth transfers.

$$WACC = r_d(1-t) \frac{D}{V} + r_e \frac{E}{V} \quad (\text{I.A.5.2})$$

where

- r_d is the cost of debt (interest rate);
- r_e is the cost of equity, assumed to be determined using the capital asset pricing model (see Chapter I.A.4);
- D/V is the proportion of debt financing and E/V is the proportion of equity financing; and
- t is the effective tax rate.

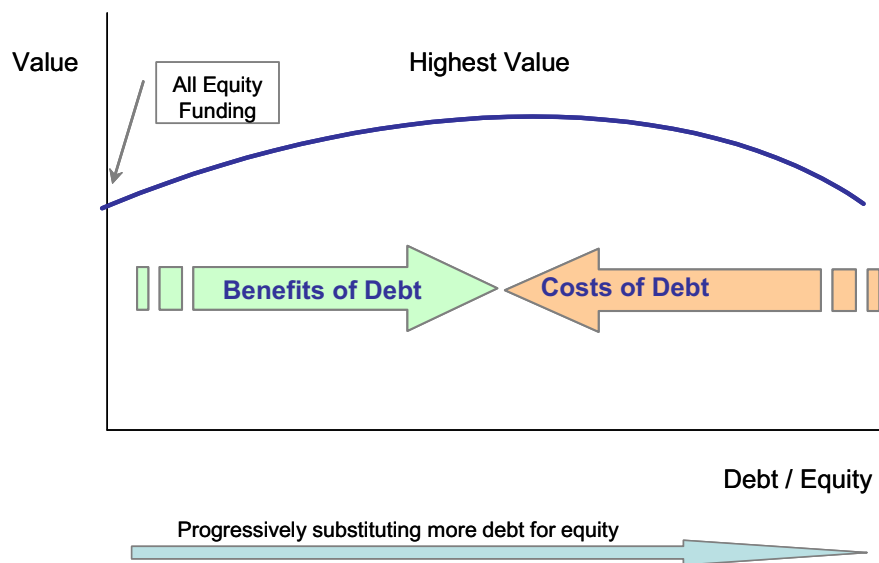
For the purposes of this chapter we will assume that the expected future cash flow can be expressed as a perpetuity and therefore its value is:

$$V = \frac{OFCF_t}{WACC} \quad (\text{I.A.5.3})$$

Our concern is with how V changes with the proportion of debt (to equity) financing, if it changes at all. The effect can be analysed by examining the impact of D/V changes on $WACC$ and/or the operating free cash flow. As we will see, both $WACC$ and cash-flow effects can arise from changing the capital structure decision.

This chapter presents both the beneficial and detrimental effects of changing the debt–equity mix to tease out the considerations in capital structure choice. The analysis of this trade-off starts with an assumption of all-equity financing and then examines how value changes as debt is progressively substituted for equity. What becomes evident is that there are benefits of debt financing that increase the value of a business up to a point where the costs begin to offset them. This point/range corresponds to an optimal capital structure, as is illustrated in Figure I.A.5.1. The shape of the curve and the optimal point/range depend upon many factors that are likely to be similar for businesses in the same industry, with some ‘uniqueness’ for a particular business. These factors or determinants are discussed in subsequent sections.

Figure I.A.5.1: Value changes with the proportion of debt financing



The chapter proceeds as follows:

- It discusses agency costs to shareholders that arise when a business appoints professional managers to operate on their behalf and, particularly, the incentives such managers have to transfer wealth from debt-holders to shareholders. This explains why we see arrangements like restrictive covenants, dividends limited to earnings, and restrictions around issuing higher-ranking debt.
- It characterises debt and equity and identifies some of the variables that distinguish different forms of debt.
- Under the heading ‘Choice of Capital Structure’ it firstly demonstrates that it is easy to be misled into thinking that debt is ‘cheap’ because the interest rate is lower than the cost of equity. Debt is not ‘cheap’. For debt to be attractive it must provide benefits other than an apparent lower interest rate. The sources of such benefits are examined next. This is followed by a discussion of the source of additional costs of debt that limits its attractiveness and therefore leads to an optimal mix of debt and equity financing.
- It finishes by presenting some guidelines for selecting the debt-equity mix and by showing what chief financial officers (CFOs) say they focus on when making this important decision.

I.A.5.2 Maximising Shareholder Value, Incentives and Agency Costs

Our understanding of the trade-offs to consider in making choices about the relative amounts of debt and equity is enhanced by understanding the incentives facing management, debt-holders and shareholders as well as the information each group has about the prospects of the firm. This section provides a backdrop to the later discussion but also enriches our understanding of why we observe the typical contents of contractual arrangements between debt-holders and the issuing firm, for example restrictive covenants in debt contracts, and why governance structures ‘overseeing’ management as so important.

I.A.5.2.1 Agency Costs

Agency costs arise when someone else makes decisions on your behalf. They arise because the agent may make decisions that are not value-maximising for the principal, that is, they are relatively suboptimal. The need for agents (managers) in corporations arises because of the benefits of specialisation, economies of scale and scope, and the associated challenges in accessing capital markets to achieve these benefits. The modern corporation brings together the following parties:

- professional managers acting as agents for shareholders (the principals) – specialists in initiating ideas, decision-making and implementation;
- board of directors – specialists in ratifying key management decisions, monitoring performance and appointing management to operate in the interests of shareholders;
- debt-holders – specialists in providing capital under contractual arrangements; and
- shareholders/residual risk bearers – specialists in providing capital and diversifying to minimise risk.

These ‘typical’ arrangements lead to agency costs of equity arising from potential conflicts of interest between management and shareholders because managers are naturally motivated to act in their own interests. Similarly, these arrangements lead to agency costs of debt arising from conflicts between the interests of debt-holders and management when they are acting in the interests of shareholders.

I.A.5.2.1.1 Agency Cost of Equity

Agency costs of equity arise because of the separation of decision-making from ownership and risk bearing. Management will be motivated by their interests rather than those of shareholders. To the extent these diverge there can be costs to shareholders. For example, management might be motivated by maximising the size of the firm and therefore their prestige and power rather than maximising the value of the firm. Maximising size can mean taking on, or retaining, value-destroying investments, over-investing in staff, expanding into unrelated areas of business that

increase size but not value.³ Mechanisms can be put in place to minimise agency costs but these cost money; for example, a board of directors is supposed to oversee management in the interests of shareholders, but such oversight, along with other monitoring systems, is costly. So agency costs not only include the opportunity cost of suboptimal decisions by management but also the costs of monitoring management to minimise these suboptimal decisions. Such mechanisms are not needed in the case of an owner-managed firm. However, owner-managed firms are not practical in large capital-hungry businesses so professional managers are employed and shares are offered more widely. This brings the benefit of skilled decision-makers but also the associated agency costs mentioned above. In a large publicly listed firm each shareholder typically owns a small portion of it and has little direct influence over it. Not surprisingly, shareholders are generally less focused in their interest.

I.A.5.2.1.2 Agency Costs of Debt

For this discussion we will assume that management does operate in the interests of shareholders, which is their primary responsibility. There are a number of ways that management, operating in the interests of shareholders, can transfer wealth from debt-holders to shareholders. In anticipation of this, debt-holders may take a number of actions to protect their interests, for example agreeing to restrictive covenant arrangements to minimise the likelihood of this happening and/or charging a higher interest rate to compensate them for expected losses. We also observe different forms of debt arising with different exposure to risk, with consequent different pricing arrangements. Understanding the agency costs of debt assists our understanding of why there are different forms of debt and different ‘protection’ arrangements.

Debt-holders could respond to the possibility of agency costs through ‘price protection’. That is, they could assume that management will act in the adverse way described and therefore charge an interest rate that covers these expected losses as well as providing the return on the investment they require. However, it is generally much more cost-effective for management to offer some form of guarantee that they will not actively pursue such activities. This line of thinking says that it is in the interests of shareholders to offer protection from agency costs of debt to save interest costs. For example, management could offer to employ an independent auditor to review the state of the company and thereby offer protection to debt-holders. Since the ultimate goal is to find the best trade-off between interest costs and monitoring costs, we can see that the incentive for audit is from the shareholders rather than from the debt-holders! This says that the benefit of the reduced cost of interest outweighs the cost of the audit.

³ Curiously many takeovers fall into this category.

What actions could management pursue to the detriment of debt-holders? What protection might they offer debt-holders to avoid higher interest costs?

Priority of dividends: Management could pay out cash earnings to shareholders, leaving insufficient to pay debt-holders. As a result, we see the requirement that interest be paid before dividends. Similarly, we see priorities defined in favour of debt-holders when a business either voluntarily or involuntarily winds up.

Sale of cash-generating assets and/or excessive dividend payments: Once management has raised debt to finance part of the business, they could liquidate the assets and pay a mammoth dividend to shareholders, leaving nothing but a shell for debt-holders. This transfers wealth from debt-holders to shareholders. We see several responses to this possibility. Firstly, we see general rules captured in company security regulations that prohibit dividends being paid out of capital (i.e. they cannot exceed earnings). Secondly, we see special resolutions required for return of capital or limits around the amount and price of share repurchases without special agreement by debt-holders (and shareholders). Thirdly, we see collateralised loans whereby debt-holders have first claim over particular assets and management agree not to sell them, and also that debt-holders have first claim over the assets in event of default or a wind-up.

Issue higher-ranking debt: Once management has raised debt at a particular interest rate they could subsequently issue new debt with a higher priority to receive interest (and principal in the event of bankruptcy). This leaves the original debt-holders bearing more risk than was originally expected and potentially uncompensated in terms of return, that is to say, there has been a transfer of wealth from one set of debt-holders to another. To protect against this possibility management can contract with the original group to define the amount of senior debt they will issue (which could be none). Why do we see debt with different seniority in the first place? Simply because it offers a broader range of risk/return securities for investors to choose from in order to establish their desired risk profile.

Increase operating risk of business: Again there is a potential for management to transfer risk from debt-holders to shareholders by increasing the operating risk of the business after the debt contract has been established. Management might signal that the debt will be used to finance assets with a particular level of operating risk, and then, once the debt is raised with an interest rate struck accordingly, might switch and use the debt to finance riskier assets. Collateralised loans which tie the debt to particular assets is one response to this situation. Another is for management to offer convertible debt whereby debt-holders can transfer the debt to equity if there is a high chance of this happening. Convertible debt can minimise these agency costs. This

is why we tend to see more convertible debt issued in industries/firms facing lots of investment opportunities and assets with changing values.

Agency costs of debt are even higher when a firm is facing default on its debt. In such a case managers are more inclined to make high-risk decisions and less inclined to invest in value-creating strategies. These and additional problems are discussed in Section I.A.5.4.3.

Agency costs help us understand the nature of debt contracts and the circumstances under which some forms of debt may be more prevalent than others. We also recognise that an all-equity firm experiences agency costs of equity but not of debt. As debt is substituted for equity to finance a firm's activities, the agency costs of debt increase and the importance of agency costs of equity decreases. We pick up this theme again in Section I.A.5.4.3.

I.A.5.2.2 Information Asymmetries

Implicit in the discussion of agency costs is an argument that management generally has more information about the firm's prospects than do shareholders, debt-holders or investors in general. We call this information asymmetry. Consequently, it might be expected that the larger the information asymmetry, the greater will be the agency costs and the greater the value investors will place on monitoring costs. For example, high-tech companies have greater difficulty attracting public debt than equity because of uncertainty about the earnings stream. This uncertainty arises in part from the challenge facing investors and analysts in being on top of the prospects of a firm when the product/service on offer is highly technical in a rapidly changing environment. However, bank debt is a different prospect. Banks generally monitor the performance of their loans closely and develop knowledge which reduces information asymmetry. Interestingly, there is evidence that the share price of firms where information asymmetry is likely to be high between investors and managers rises upon the announcement of the issue of bank debt.

Information asymmetries explain a number of market reactions to management actions that otherwise seem odd. For example, there is strong evidence that when a firm issues additional equity the share price falls. Since equity is used to finance value-creating investments it at first seems odd that the share price falls. One explanation that seems to fit the facts pretty well is that managers know more about the firm's prospects than anyone else. Further, they issue equity when they think the equity value is high. Investors will therefore interpret the issue of equity as bad rather than good news, that is, they interpret the issue of shares as a signal that the shares are overvalued in the minds of those who know best. This leads to a downward revision in the price of the shares!

We revisit this notion when providing guidance about what proportion of debt and equity a firm should issue.

I.A.5.3 Characteristics of Debt and Equity

Debt and equity securities differ principally in the contracted sharing of business risk. Debt securities are usually contracts for an agreed interest rate (or premium over a benchmark rate) and have a defined horizon. Consequently the variability (or co-variability) of the cash-flow stream to debt-holders is relatively low – much lower than that of the operating earnings stream.

Shareholders, on the other hand, contract for residual earnings. They receive what is left after all other income claims of contracting parties to the firm are met. Further, the time horizon of the shares is infinite, although they can be traded. Consequently, the variability (and co-variability) of the cash-flow stream that shareholders can expect is generally much higher than the underlying operating earnings stream, that is, it is much riskier than debt and riskier than the earnings stream if the firm issues debt at lower risk than the underlying earnings stream.

While these are high-level distinctions between debt and equity, there are many forms of debt and even equity. Debt instruments issued by a business vary according to a number of features which influence the risk profile (Table I.A.5.1).

These different forms of debt will have different pricing and yields. The range helps ‘complete the market’ by providing a wide choice to investors with different risk profiles and provides a choice to the issuer to help match particular circumstances. For example, convertible debt is attractive to high-growth firms that need to conserve cash for reinvestment because the interest rate is generally lower than other forms of debt. It is attractive to investors who are looking to share in the upside performance of the business while being ‘protected’ to some extent on the downside. Readers interested in a more detailed discussion of the way debt can be matched to particular business circumstances should refer to Barclay and Smith (1999) and Barclay *et al.* (1995).

Note that there are a number of financing instruments that are really a mix of debt and equity. Preference shares, for example, have the characteristics of debt because, like interest, they generally offer a fixed percentage (of face value) dividend yet the dividends are treated like equity for tax purposes, that is, are not deductible. Convertible notes are clearly a hybrid with debt characteristics until converted, when they become equity capital. For this reason, regulatory bank capital has been carefully defined to be of the character of equity in Tier 1 (ordinary shares and

retained earnings but excluding cumulative preference shares), with some instruments with equity characteristics treated as Tier 2 (supplementary equity capital, e.g. hybrids, subordinated debt).

Table I.A.5.1: Characteristics of debt instruments

Feature	Explanation
Maturity	Refers to the time when the debt must be repaid or rolled over. Ideally issuing firms have a mix of maturity of debt to roughly match the maturity of the assets the debt is funding.
Balloon	Calls for one repayment of both interest and principal at maturity date.
Coupon	Calls for payment of interest to the holder of the coupon at the time the interest is due. Generally preferred by investors seeking a steady income stream.
Seniority	Refers to the order of the claim on the income/assets of the firm. More senior debt is less risky to the investor and will have a commensurately lower interest rate.
Callable	The issuer can ‘call in’ the debt, i.e. repay it within a defined period. Gives some flexibility to the issuer therefore has a higher interest rate than non-callable.
Puttable	Opposite to callable. Enables the holder to ‘sell’ the debt back the issuing business with a defined period
Convertible	Enables the debt-holder to convert the debt to equity during a defined period. Generally has a lower explicit interest rate than non-convertible debt but confers additional value to the holder because of the potential upside associated with the equity component. Useful for firms with positive growth opportunities.
Collateralised	The debt has a claim over specific assets. In the event of default, the debt-holders may take over control of the asset to recover the monies owed to them.

I.A.5.4 Choice of Capital Structure

To understand the capital structure choice we commence with a relatively simple case and build from there. In the first case we show that there is no real advantage to having or not having debt financing, despite the fact that the rate of return debt-holders require (the interest rate) is lower than the rate of return shareholders want, that is, it seems cheaper! This is a key point because, firstly, it initially seems counterintuitive, and secondly, it forces us to look further to better understand why debt might be attractive.

We look further for advantages in Section I.A.5.4.2 and we find some real advantages. However, nothing is free! There are disadvantages of increasing debt funding, and we identify these in

Section I.A.5.4.3. As we increase the proportion of debt funding we increase these disadvantages, and eventually these can be expected to offset the advantages. So there will be a point at which we should stop increasing the proportion of debt funding. The exact point at which this happens is easy to talk about but hard to define in practice, so in Section I.A.5.5 we look at what CFOs say they look at. We see how this fits in with our discussion so far, and then extract some guidelines for selecting an appropriate value-maximising financing mix.

I.A.5.4.1 Do not think debt is attractive because the interest rate is lower than the cost of equity!

Imagine you manage a health insurance company which is looking at adding a financial planning business to the current customer offer. This seems a good idea because your existing customers have demonstrated concern about the financial effects of ill-health, already trust your product, use your distribution channels, and you already have a funds management business to invest the health insurance premium income. Further, you have access to some customer financial details and could commence by offering the financial planning service to the higher-wealth group on the assumption that they are the most likely to buy.

The business needs investment to get it up and running. It needs to buy/develop specialist software, training courses, supporting research services, etc. This investment is \$1 million. Once the business is up and running you estimate it will earn a pre-tax cash-flow return of \$150,000 per year – a rate of return on investment (ROI) of 15%. (We will assume for the sake of simplicity that there is no corporate tax.) Supposing that you can finance this investment *either* by equity at 12% required rate of return (estimated from the capital asset pricing model using a risk-free rate of 6%, market risk premium of 6% and beta of equity of 1), *or* by debt at an interest rate of 8%, which will you choose?

Looks easy, doesn't it – debt looks much cheaper than equity!

Not only does it look cheaper but your investment banking advisers say that the debt will do wonders for the return on equity (ROE). Take a look at Table I.A.5.2, prepared by the bank. It shows the return on equity for the business under three different funding arrangements from all equity to 25% equity and 75% debt.

Table I.A.5.2: Impact of increasing leverage on return on equity

	Financing Method		
	All Equity	50% Debt 50% Equity	25% Equity 75% Debt
Equity Investment	\$1,000,000	\$500,000	\$250,000
Debt Investment		\$500,000	\$750,000
Total Investment	\$1,000,000	\$1,000,000	\$1,000,000
Debt / Equity Ratio	0	1	3
EBIT ⁴	\$150,000	\$150,000	\$150,000
Interest	0	40,000	60,000
Net Income	\$150,000	\$110,000	\$90,000
Return on Equity	15%	22%	36%

ROE grows as we have more debt. The logic seems OK, because the business is earning 15% on each dollar of investment and the debt-holders only want 8% for each bit they finance, leaving much more than 15% to the shareholder-financed proportion (i.e. they get 15% on each dollar they finance plus the difference between 15% and 8% on the debt-financed portion). For those who like formulae, the relationship is:

$$ROE = ROI + (ROI - r_d) \frac{D}{E}, \quad \text{I.A.5.4}$$

that is, the return on equity is a positive linear function of the debt to equity ratio (provided ROI is greater than the interest rate r_d).

What is wrong with this apparently rosy picture? The analysis assumes that \$150,000 will be earned from the investment. However, this is only an estimate (or mean) from a range of possible outcomes. Look at what happens to ROE if we look at possible outcomes both higher and lower than \$150,000 (Table I.A.5.3). ROE looks great if the outcome is better than \$150,000 but there is a real downside – it looks much worse when the interest bill is not covered.

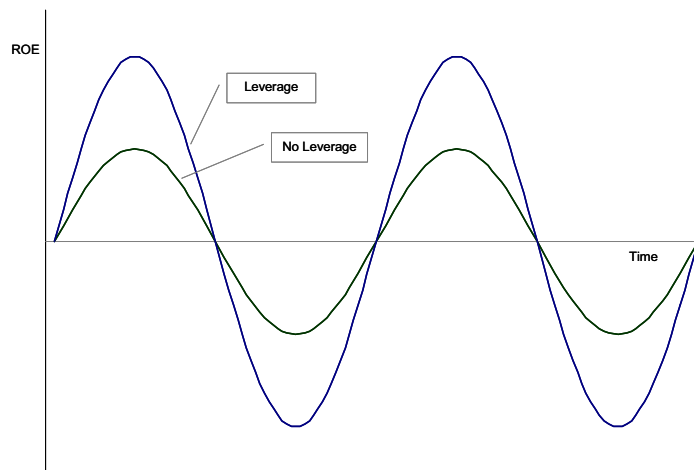
⁴ EBIT means Earnings Before Interest and Tax

Table I.A.5.3: How ROE varies if earnings vary from the mean

Financing Mix	Return on Equity for different EBIT levels (\$'000)						
	0	50	100	150	200	250	300
All Equity	0%	5%	10%	15%	20%	25%	30%
50% Debt, 50% Equity	-8%	2%	12%	22%	32%	42%	52%
75% Debt, 25% Equity	-24%	-4%	16%	36%	56%	76%	96%

Further suppose the earnings change over time, each outcome being drawn from the probability distribution of possible outcomes but appearing cyclically (to demonstrate the higher variance in return to shareholders) and the pattern is captured in Figure I.A.5.2. Notice how the variability in ROE is greater than the variability in ROI (no leverage), that is, risk is increased. So the alleged benefit of increased leverage can be a double-edged sword. The benefit to shareholders is achieved at the expense of increased risk. Not only does increasing the proportion of debt make the earnings available to shareholders more variable (increases variance) but it also increases risk (variability and beta).

Figure I.A.5.2: ROE is magnified by leverage



Increasing equity risk means that shareholders will require a higher rate of return, thus there are two costs associated with increasing leverage:

- the explicit interest cost of debt;
- the implicit rise in the cost of equity in response to the higher risk they face.

These two effects offset each other exactly under certain conditions so that there is no overall benefit of the apparently cheaper debt!

This should not be all that surprising when we recognise that the overall risk of a business is determined by its investment decision, not by the way it is financed. If the financial planning business is financed entirely by equity then clearly all the risk is borne by shareholders. If it is financed by some combination of debt and equity then that business risk is being shared, less by the debt-holders and more by the shareholders. However, this does not change the nature/risk of the business's cash-flow earning stream, just the distribution of it. The expected business earnings are \$150,000 regardless of how the business is financed. Further, the probability of possible business earnings is not changed either. The \$150,000 is a function of interaction with customers and the cost of those interactions, not the interest bill. So if the underlying risk that is shared and the underlying cash flow that is also shared between the debt-holders and the shareholders do not change because of the financing mix then neither will the value. Thus capital structure will not matter.

This very important point can be made another way. Suppose you were planning to start the business financing it yourself by having equity issued to you. You would assess the risk of the business and determine the rate or return you would require on this equity. Suppose this is 12%. If the earnings stream was a perpetuity then you would value the business at follows:

$$V = \frac{150,000}{0.12} = \$1,250,000$$

Suppose instead you decided to finance the business yourself partly by equity and partly by debt. What rate of return would you require? Overall there is no reason for the required return to be anything other than 12%. Nothing about the risk of the business has changed. The only difference is that you have taken debt and equity securities instead of all equity. You would require a lower rate of return than 12% on debt because it has a constant interest stream and a higher rate of return than 12% on the equity because it now has a subordinate claim to the debt, but the weighted average of the required rates of return should still be 12%. Consequently, there is no reason for the overall value of the business to change from \$1,250,000 and there can be no value benefit from the different capital structure.

The circumstances that must prevail for this conclusion are reasonably stringent but the conclusion that the way a business is financed does not matter is valid under these circumstances. It has demonstrated a very important point – that there are two costs of debt, an explicit and an implicit cost (the increase in the cost of equity due to the additional risk), and the latter is often ignored. The circumstances that must prevail revolve around there being no special tax benefits of debt or adverse consequences of default on debt.

Before leaving this section, one of the assumptions underlying the analysis is that debt is priced ‘correctly’ with respect to risk. Cheap debt is really mispriced debt. Some argue that deposit funds are a cheaper source of funds than wholesale funds for banks. This can be true once all costs of deposit funds are accounted for. That is, a large infrastructure is required to attract and manage the low-interest-paying source of funds. Until these costs have been covered by the quantum of funds times the ‘saving’ on interest relative to wholesale funds it is hard to say that deposit funds are a cheap source of funds. Nevertheless, once the infrastructure is in place, the challenge for a bank is to use this source as much as possible.

To better understand the factors that make the capital decision relevant, we now look at changes in the special circumstances of this section by introducing both benefits and costs of adding debt to an otherwise all-equity-financed business.

I.A.5.4.2 Debt can be attractive

There are a number of circumstances when increasing the proportion of debt financing will either improve the expected cash flow or reduce the cost of capital, thereby increasing the value of the business.

I.A.5.4.2.1 Differential treatment of payments to debt-holders and shareholders

Classical taxation systems, such as that currently operating in the USA, treat payments of interest to debt-holders more favourably than payments to shareholders. Payment of interest is deductible, whereas payment of dividends is not. Consequently, financing with debt offers an advantage that financing with equity does not. The total cash outflow to the tax authorities will be lower, the greater the debt employed. This is, in fact, a powerful incentive to use debt as the value of the firm will rise as the funding mix is changed from all equity to more and more debt.

Illustration

The value of the health insurance business in the last section was \$1,250,000, valued as follows:

$$\begin{aligned} V &= OFCF / WACC \\ &= 150,000 / 0.12 \\ &= \$1,250,000. \end{aligned}$$

If corporate tax is now introduced at 30% (t_c) the value will fall because less earnings are available for distribution. Or put another way, the government has taken a portion of the value of the firm to meet broader society goals. In this case the relevant operating free cash flow will be

after tax and becomes $OFCF(1 - t_c)$ or $150,000 \times 0.7 = \$105,000$. The value of the business becomes:

$$\begin{aligned}V &= OFCF(1 - t_c) / WACC \\ &= 105,000 / 0.12 \\ &= \$875,000.\end{aligned}$$

What happens if the business is financed in part by debt? Suppose \$500,000 of debt is raised at 8%. Here the business gets relief on its tax bill because of the interest payment. Tax will be \$12,000 lower each year:

$$\begin{aligned}\text{Tax relief p.a.} &= \text{Interest} \times \text{tax rate} \\ &= \text{Debt} \times \text{interest rate} \times \text{tax rate} \\ &= 500,000 \times 0.08 \times 0.3 \\ &= 12,000.\end{aligned}$$

Since this will be received each year, and its risk is a function of the risk of the debt, the present value (PV) of this benefit in perpetuity will be

$$\begin{aligned}\text{PV of tax relief} &= (\text{Debt} \times \text{interest rate} \times \text{tax rate}) / \text{interest rate} \\ &= 12,000 / 0.08 \\ &= \$150,000,\end{aligned}$$

and the value of the business will be higher by this amount, that is,

$$\begin{aligned}V &= \text{value before plus PV of tax relief} \\ &= 875,000 + 150,000 \\ &= \$1,025,000.\end{aligned}$$

The implication of this analysis is that the greater the debt employed, the bigger the tax shield and the higher the value of the business!

In this illustration, all the benefit flowing from the lower tax payment has been treated as an increase in the operating free cash flow after tax, relative to the all-equity case. The benefit of lower tax could also have been calculated by keeping the after tax operating free cash flow at \$105,000 and reflecting the benefit as a lower $WACC$.

The tax benefit associated with debt arises because interest payments to debt-holders are tax-deductible but dividend payments to shareholders are not. Imputation tax systems, such as that

operating in Australia, work to reduce this bias. The bias is not fully removed, however, as overseas investors do not receive the tax rebate on tax paid on dividends which is available to Australian residents. Consequently, debt is less attractive from this perspective in Australia than in the USA.

I.A.5.4.2.2 Greater Flexibility

Retaining earnings, rather than paying dividends, is the lowest transaction cost method of raising equity. In rapidly growing businesses, however, the funds available from this source are often inadequate. Further, investment opportunities generally do not offer themselves concurrently with the availability of cash earnings. Raising new equity, on the other hand, is quite expensive and time-consuming.

Debt can often be raised relatively quickly and when needed with little notice – for example, stand-by facilities provide funds with little time delay.

Debt finance gives greater flexibility so that value-creating opportunities can be exploited. Undue reliance on equity finance might mean that value-creating opportunities are passed over, causing a relative loss in shareholder value.

I.A.5.4.2.3 Monitoring 'improves' performance and reduces the negative aspect of information asymmetry

An interesting empirical finding noted earlier is that the value of a firm rises when it announces bank-issued debt. One explanation for this is that banks are good at watching the value of their debt. They require frequent reporting from the firm so that they can carefully monitor its prospects, the quality of its strategy and the performance of management. The banks therefore serve to reduce the negative impact of information asymmetry between investors and management; investors have greater confidence that their interests are partially protected as the bank protects its interests. Consequently, one positive impact of debt is the increased monitoring of the prospects of the firm by interested parties, which can lead to an increase in shareholder value.

I.A.5.4.2.4 Debt enforces a discipline of paying out operating earnings

Another potential benefit of having debt financing is that it reduces the opportunity for management to invest earnings in bad investments (negative net present value (NPV) investments). By being committed to meet interest payments, management has less unused funds to divert to poor decisions. With high interest payments relative to income, there is less uncommitted cash to invest. Consequently, with high debt payouts, if management wants to undertake a substantive investment opportunity it has to go to the market to request the capital.

The investment opportunity is subjected to much more careful scrutiny than if funded through retained earnings.

This has greatest importance for mature businesses that generate large amounts of cash and have few positive NPV investment opportunities. Here we might expect to see a higher proportion of debt financing in mature industries where management face the temptation of investing for size rather than value!

It is also argued that the high level of debt and consequent interest payments sharpen management's attention in leveraged buyouts and management buyouts. These transactions are typified by a substantial proportion of debt financing providing a challenge for management to turn the business around and pay down the debt to a level sustainable in the longer term.

I.A.5.4.2.5 Debt financing avoids negative signals about management's view of the value of equity

It was noted earlier that there is well-documented evidence that, on average, the value of equity falls when a firm issues new shares. This is attributed to management having more knowledge about the firm's prospects than do investors. Managers tend to issue equity when they perceive it to be over-valued and the market responds accordingly with a downgrade of share price. Use of debt rather than equity minimises this impact. Myers and Majluf (1984) argue that a firm should use internal sources of funds first, then debt, then hybrid, and finally equity for this reason.

I.A.5.4.3 Debt can also be unattractive

There are many reasons why debt can be attractive, but it can also be unattractive. Increasing the proportion of debt funding can decrease the value of the firm because:

- it increases investors' exposure to a loss of value if bankruptcy occurs (direct bankruptcy costs);
- it also exposes investors to the loss of value if financial distress arises due to challenges in meeting high interest payments (indirect bankruptcy costs); and
- it encourages management to pass up value-decreasing investment opportunities where the benefit flows only to debt-holders, and it encourages high-risk investments – 'betting the bank' (agency costs).

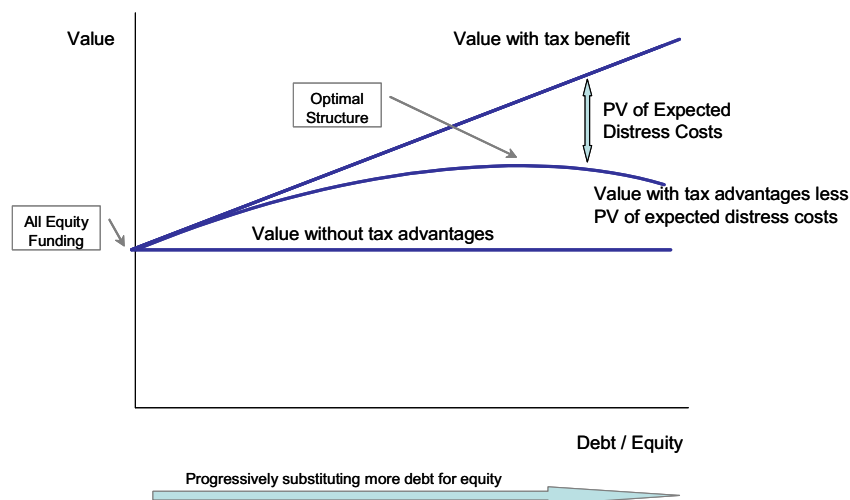
I.A.5.4.3.1 Exposure to bankruptcy costs

Bankruptcy can be expensive. Court and legal costs use up the resources of a firm which otherwise belong to the debt-holders and shareholders. Such costs are only incurred because of debt, and the more debt the higher the probability of incurring these costs. Thus there is a trade-off occurring as proportionately more debt financing is employed. On one hand, increasing the

proportion of debt leads to the benefits described above, but on the other hand it exposes the firm to a greater probability of incurring bankruptcy costs. Thus there ought to be an optimal proportion of debt occurring when the marginal benefit is offset by the marginal cost, as indicated in Figure I.A.5.3.

There have been studies examining whether this trade-off can explain the proportions of debt financing employed by firms.⁵ Unfortunately the size of bankruptcy costs is small relative to the tax benefit of debt and cannot explain the debt–equity mixes we observe. That is, the size of these costs is so relatively small that debt–equity mixes ought to be of the order of 90–100%. As we typically observe much lower levels of debt, this led researchers to focus on a broader view of costs incurred by having debt in capital structure, viz. distress costs.

Figure I.A.5.3: Trade-off of benefit and costs of increasing the proportion of debt funding



I.A.5.4.3.2 Exposure to financial distress costs

Indirect financial distress costs arise because customers and investors are frightened by the prospect that a company might cease to exist. As a result sales can fall dramatically, leading to loss of cash flow and even greater difficulty in meeting interest payments. Opler *et al.* (1997) cite two examples of this. Chrysler’s share of the motor car market dropped from 30% to 13% when consumers learned that it was in trouble in 1979. Caldor (a US retailer) lost the support of its

⁵ Warner (1977) measured the direct costs of bankruptcy for a number of railway companies undergoing bankruptcy between 1933 and 1955. These costs averaged 1% of the market value of the firm for 7 years before bankruptcy and rose to 5.3% just before bankruptcy. Pham and Chow (1989) found them to average 2.5% and 3.5% of firm value the year before and the year of bankruptcy.

trade creditors when its sales fell after another retailer (Bradlee's) had filed for Chapter 11 bankruptcy protection. This 'reflected' outcome led to Caldor facing bankruptcy itself. In Australia, customers deserted OneTel, a telecommunications company, when news broke that it was having cash-flow difficulties. This added to the company's woes and it ended up being wound up.

In addition to loss of sales, the cash-flow difficulties can lead to cutting costs such as essential maintenance and otherwise value-creating activities, for example research and development, leading to even further difficulties.

Pham and Chow (1989) found these costs to be non-trivial for their sample of firms. Both direct and indirect costs were estimated to be over 20% of the value of the firm.

The distress costs associated with banks are expected to be very high and the cost to society is also seen to be very high. If a bank fell into financial difficulty there is an expectation that depositors would rush to withdraw funds, thereby creating a liquidity crisis. Consequently, this is probably a significant driver of capital structure choice for them. Further, because of the potential contagion effect, regulatory authorities in many countries adhere to the Basle Accord (see the introductory Chapter III.0 in Section III).

I.A.5.4.3.3 Agency costs

Agency costs of debt can be very high when a firm is in technical default of its debt obligations, that is, it has not yet defaulted on its debt obligations but default is most likely when a debt-servicing payment falls due. Note that these circumstances do not arise when a firm has no debt, thus these are clearly a cost of debt financing. Two examples of management behaviour that make sense when a firm is experiencing financial distress are explained below.

Bet the bank when under duress: If equity is technically of zero value, the value of debt is less than its face value but the firm has not yet liquidated, there is an incentive to 'gamble'. Rather than pay out the remaining resources to the debt-holders, managers might invest in high-risk activities. If the gamble pays off, equity-holders can then payout the debt and have a positive value from their investment. Note that here the managers (acting on behalf of shareholders) have nothing to lose. Without taking on the gamble, the equity has zero value. If the gamble does not pay off, the debt-holders will bear the consequences. Offering to restrict these activities through covenants can decrease the flexibility management has and can lead to loss of positive NPV investments.

Under-invest when benefits flow to debt-holders: Consider a firm that finds itself in the circumstances described above: the value of debt is below the principal amount it has to pay out and the value of equity is zero. In these circumstances it may forgo undertaking value-increasing investments, and this opportunity cost is an agency cost of debt. For example, suppose there is a payment on debt of \$100,000 due in 6 months' time but the current value of the firm is \$90,000, so that debt-holders face a loss of \$10,000. In these circumstances the value of equity is zero as shareholders will get nothing from the value of the business. Suppose, further, that there is an investment opportunity that has a positive NPV of \$9,000. In this case, there is no incentive for management, operating in the interest of shareholders, to undertake the value-increasing investment. Shareholders would be asked to contribute more capital to help fund the investment project yet they do not share in the positive NPV – it all goes to make up the \$10,000 deficiency in debt value. Consequently, value-enhancing investment opportunities can be forgone when a firm is facing default. Some, among them Myers (1977), argue that this is a substantive agency cost of debt.

I.A.5.4.4 Thus choose the point where disadvantages offset advantages

Figure I.A.5.1 captures the two opposing forces at work as a firm increases the proportion of debt funding of its assets. Note how the value of a firm can increase as debt is increasingly substituted for equity due to the many benefits described in Section I.A.5.4.2. As the proportion of debt increases, so does the probability of incurring bankruptcy and distress costs. A point, or range, is reached whereby the marginal disadvantage is greater than the marginal advantage and this will define an optimal capital structure.

I.A.5.5 Making the Capital Structure Decision

The discussion so far has identified the capital structure decision to be a trade off of the advantages and disadvantages of increasing the proportion of debt financing. The choice for a particular company will depend upon the particular circumstances it faces. This section discusses how particular circumstances a firm faces might influence its choice and also discusses what CFOs have said they consider when making capital structure decisions.

I.A.5.5.1 Guidelines

We can expect that firms will choose a capital structure that best fits its particular circumstances. Some guidelines around some of the capital structure choice considerations are summarised in Table I.A.5.4. Note that the table highlights the link between distress costs and variability of asset values. Chapter III will show that for a bank, the amount of capital required is directly linked to value-at-risk. This is a measure derived from the variability of the bank's assets.

Banks have very high leverage compared with their industrial counterparts. This is in part because they will generally have less volatile earnings streams. For example, the earnings stream from lending is quite well defined and stable, leading to low volatility. It is also subject to careful credit risk scrutiny which minimises the variability. Further, banks tend to have relatively few other substantive tax shields (depreciation, research and development) enabling greater access to the tax advantages of debt. Borrowing can further be supported by deposit insurance programmes that partially mitigate the cost of distress.

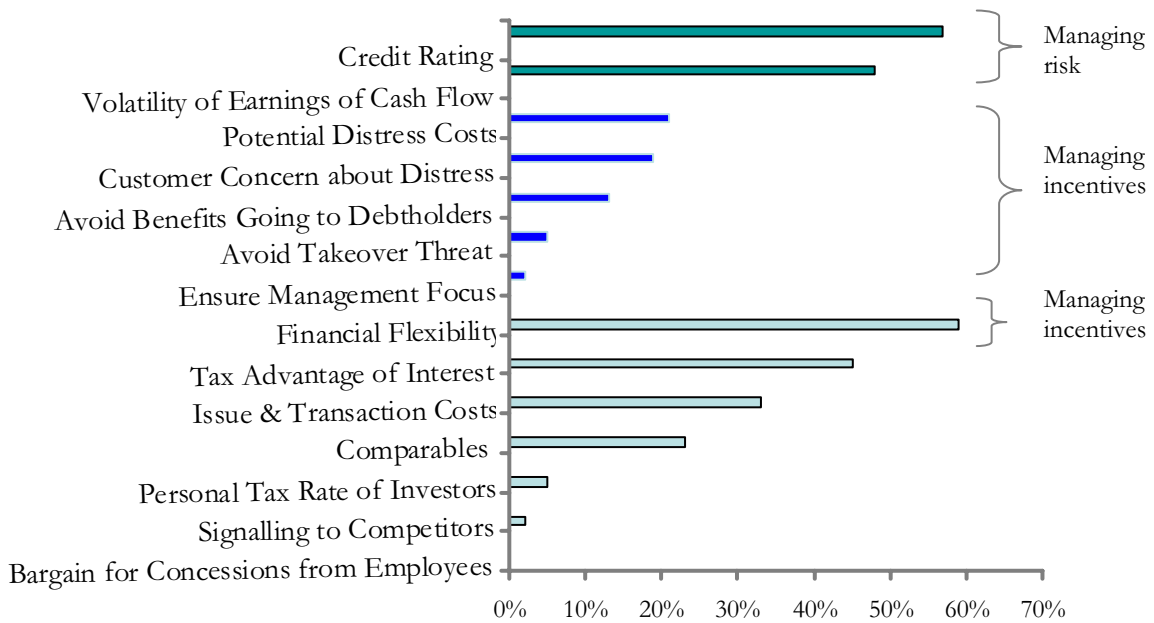
I.A.5.5.2 What do CFOs say they consider when making a capital structure choice?

A recent survey by Graham and Harvey (2001) asked US chief financial officers to describe the considerations that influenced the choice of capital structure. As can be seen from the extract from that survey (Figure I.A.5.4), the considerations are consistent with the discussion in the chapter.

Table I.A.5.4: Guidelines for selecting a capital structure

Consideration	Guideline	Reasoning
Tax Benefit	<ul style="list-style-type: none"> Firms with other tax shields will have less debt. 	<ul style="list-style-type: none"> High depreciation, research and development and other tax deductions reduce the ‘room’ for interest to be claimed.
Distress Costs	<ul style="list-style-type: none"> Businesses with highly volatile earnings streams or those whose assets are highly volatile will have less debt. Larger firms will have more debt. 	<ul style="list-style-type: none"> High volatility means greater risk of incurring bankruptcy costs, therefore less debt capacity. The bankruptcy and distress costs have a large fixed cost component. These become relatively less important the larger is the firm.
Agency costs	<ul style="list-style-type: none"> Firms with a higher proportion of intangible assets and growth options will have less debt. 	<ul style="list-style-type: none"> This generally makes it harder to sell collateral and it also generally means more opportunity to change risk and asset profile, thus greater agency costs of debt.
Flexibility	<ul style="list-style-type: none"> Firms with more growth opportunities will have more unused debt capacity. 	<ul style="list-style-type: none"> Growth opportunities generally mean a need to act quickly and flexibly, thus a demand for standby-type facilities.

Figure I.A.5.4: Factors influencing debt–equity mix



Source: J R Graham & C Harvey "The Theory and Practice of Corporate Finance" Working Paper

Interestingly, flexibility and credit ratings rate highly. Many CFOs will choose a desired credit rating then choose a capital structure that enables the firm to meet the criteria for that credit rating. While there are qualitative overlays that rating agencies apply to financial ratios, financial ratios can guide the likely rating of a credit organisation. Table I.A.5.5 indicates the median ratio for firms within Standard and Poor's credit ratings. Thus, given forecast earnings before interest and taxes, a firm can try different capital structures and therefore different financial ratios to 'fit' within a desired credit rating and therefore interest cost.

For example, if the health insurance company considered in Section I.A.5.4.2 employs \$500,000 in debt, its interest expense at 8% p.a. will be \$40,000. With expected earnings of \$150,000, interest cover is earnings/interest or 3.75. This places it in the BB to BBB range. To be rated AA on this measure the EBIT interest cover ratio would have to be 9.2⁶. This sets the interest expense at \$16,300 (\$150,000/9.2) which corresponds to debt of \$293,750. Sensitivity analysis based on the volatility of earnings can further inform this choice.

Analysis of this type is consistent with the drivers of capital structure as discussed in this chapter. Research has shown that there is a relationship between credit rating and probability of

⁶ EBITDA is Earnings Before Interest, Taxes, Depreciation and Amortisation expenses i.e. a proxy for cash flow

bankruptcy. Thus high-rating firms will generally have lower proportions of debt and a lower probability of bankruptcy, as the discussion predicts.

Table I.A.5.5 Median ratio values for different credit ratings⁷

Industrial long-term debt

Three year medians (1996 to 1998)

	AAA	AA	A	BBB	BB	B	CC
EBIT interest cover	12.9	9.2	7.2	4.1	2.5	1.2	-0.9
EBITDA interest cover	18.7	14.0	10.0	6.3	3.9	2.3	0.2
Funds Flow/Total Debt	89.7	67.0	49.5	32.2	20.1	10.5	7.4
Free operating cash flow / total debt	40.5	21.6	17.4	6.3	1.0	-4.0	-25.4
Return on Capital	30.6	25.1	19.6	15.4	12.6	9.2	-8.8
Operating Income / sales	30.9	25.2	17.9	15.8	14.4	11.2	5.0
LT Debt / capital	21.4	29.3	33.3	40.8	55.3	68.8	71.5
Total Debt / Capital	31.8	37.0	39.2	46.4	58.3	71.4	79.4

Source: *Standard and Poor's Credit Week*, 28 July 1999

I.A.5.6 Conclusion

How should a business be funded? How much debt? How much equity? Does it matter? Our discussion shows that the choice of capital structure does affect the value of the business and therefore it is an important decision area for managers and shareholders. The effect is both direct, for example though the taxes payments 'saved', and indirect, through the impact on managers' motives and decision-making. Exactly what the best funding mix is for a particular business is a little imprecise, however there are a number of key guidelines that we can provide to help management make shareholder value-creating decisions in this area.

A key determinant of capital structure choice for a business is the risk or variability of the operating earnings stream. Generally, the higher this risk the less debt can be supported, essentially due to the exposure to distress and bankruptcy costs. Many firms will select a desired credit rating or probability of distress based on the volatility of the earnings stream and use this to establish the interest bill, or amount of debt that can be supported.

Financial institutions are a special class of company in the sense that there are high costs of distress associated with them. As a consequence, the choice of capital structure is generally directed at keeping the exposure to bankruptcy and distress costs at an 'acceptable level'.

⁷ Source: *Standard and Poor's Credit Week*, 28 July 1999

References

- Barclay, M. and Smith, C W (1999) The capital structure puzzle: another look at the evidence. *Journal of Applied Corporate Finance*, 12(1), pp. 8–20.
- Barclay, M, Smith, C W and Watts, R L (1995) The determinants of corporate leverage and dividend policies. *Journal of Applied Corporate Finance*., 7(4), pp. 4–19.
- Graham, J R and Harvey, C R (2001) Theory and practice of corporate finance – evidence from the field. *Journal of Financial Economics*, 60, pp. 87–243
- Myers, S C (1977) Determinants of corporate borrowing. *Journal of Financial Economics*, 5(2), pp. 147–175.
- Myers, S C and Majluf, N S (1984) Corporate financing and investment decisions when firms have information that investors do not have. *Journal of Financial Economics*, 12, pp. 187–221.
- Opler, T, Titman, S and Saron, M (1997) Corporate liability management: designing capital structure to create shareholder value. *Journal of Applied Corporate Finance*, 10(1).
- Pham, T and Chow, D (1989) Some estimates of direct and indirect bankruptcy costs in Australia: September 1978 – May 1983. *Australian Journal of Management*, 14(1), pp. 75–96
- Warner, J B (1977) Bankruptcy costs: some evidence. *Journal of Finance*, 32(2), pp. 337–348.

I.A.6 The Term Structure of Interest Rates

Deborah Cernauskas and Elias Demetriades¹

Interest rates are at the heart of modern finance. They are a focal point for choosing valuation models, developing trading strategies and measuring and managing risk. Given the dynamic nature of all these activities, understanding how interest rates are calculated and how they change over time is of critical importance. These two issues are the focus of this chapter.

A fundamental concept in finance is that money has a time value that results from investment opportunities. A fixed income investment made today for a specified term will result in a payoff or future value that is dependent on the compounding method employed. Section I.A.6.1 covers different compounding methods used to calculate the future value of an investment. Section I.A.6.2 recognizes that interest rates, paid or charged for money, depend on the length of the investment term and explains the concept of the term structure of interest rates. Section I.A.6.3 illustrates common shapes of the yield curve. Section I.A.6.4 illustrates two important derivations of the yield curve: spot and forward rates. Section I.A.6.5 concludes with a summary of the common theories that explain the shape of the yield curve.

I.A.6.1 Compounding Methods

Receiving a dollar today is not equivalent to receiving a dollar in a year's time. A dollar received today can be invested at the prevailing interest rate for a year, resulting in an amount greater than one dollar. The interest rate represents the price paid to use money for a period of time and is commonly referred to as the *time value* of money. The future value of an amount invested today will depend upon how the interest is calculated. There are two methods of calculating interest: *discrete* compounding, which includes simple and compound interest calculations; and *continuous* compounding.

I.A.6.1.1 Continuous versus Discrete Compounding

When an investment earns simple interest, interest is earned only on the amount invested or 'principal'. The future value (FV) of an investment earning simple interest is given by the formula:

$$FV = P(1 + rt) \quad (\text{I.A.6.1})$$

¹ Deborah Cernauskas Ph.D. is Visiting Professor of Finance, Department of Finance, Northern Illinois University. Elias Demetriades is Adjunct Professor of Finance, Illinois Institute of Technology.

where P is the principal, r is the annual interest rate and t is the length of time of the investment in years. Simple interest is typically used when there is only a single time period.

Consider a principal of \$100 invested for one year earning 5% simple interest. At the end of the one-year period, the investment is worth

$$\$100(1.05 \times 1) = \$105$$

The same investment over a three-year period is worth

$$\$100(1 + 0.05 \times 3) = \$115$$

Time	0	1	2	3
	----- ----- -----			
Value	\$100	\$105	\$110	\$115

The investment earns \$5 per year in interest, and at the end of the three-year period the investment has earned \$15. Many investments pay compound interest instead of simple interest. An investment that pays compound interest pays interest on both the principal and previous interest payments. The payment of interest on interest is a significant difference between simple and compound interest. The future value of the principal amount earning compound interest is given by the formula:

$$FV = P(1 + r)^n \tag{I.A.6.2}$$

where P represents the amount invested, r is the per period interest rate, and n is the number of periods.

Consider an investment of \$100 invested for three years earning 5% per annum compounded annually. At the end of the three-year period, the investment is worth:

$$FV = \$100(1 + 0.05)^3 = \$115.7625$$

Time	0	1	2	3
	----- ----- -----			
Value	\$100	\$105	\$110.25	\$115.7625

An investment of \$100 earning simple interest has a future value of \$115 after three years. The same investment will earn an additional \$0.7625 when interest is computed using annual compounding

In certain situations, it is appropriate to consider interest earned through continuous compounding. Investments earning continuously compounded interest earn interest so frequently that the time period between interest-rate calculations approaches zero. In practice, no one compounds interest continuously but it is used extensively for pricing options, forwards and other derivatives. The future value of an investment earning continuously compounded interest is given by the formula:

$$FV = Pe^{rt} \quad (\text{I.A.6.3})$$

P represents the amount invested, r is the annual interest rate, t is the number of years of the investment, which can be a fractional amount, and e is the number 2.7182818285 (to 10 decimal places).

Example I.A.6.1

Consider an investor who deposits an amount of \$5000 in a bank account paying an annual interest rate of 4.2%, compounded continuously. What is the balance after two and a half years?

To find the future value of the account, use the continuous compound interest formula (I.A.6.3) with $P = 5,000$, $r = 0.042$, and $t = 2.5$.

$$FV = 5000e^{0.042 \times 2.5} \approx \$5553.55$$

The balance at the end of 2.5 years is approximately \$5553.55.

I.A.6.1.2 Annual Compounding versus More Regular Compounding

In the last section, the discrete compounding period was assumed to be one year. Sometimes interest is paid more frequently. Interest on savings deposits may be paid semi-annually, quarterly or monthly. Interest on car loans and mortgages is typically calculated monthly. The general compound interest formula is given by:

$$FV = P \left(1 + \frac{r}{n} \right)^{nt} \quad (\text{I.A.6.4})$$

where P is the principal invested, r is the nominal interest rate per annum, n is the number of compounding periods per year, and t is the term of the investment in years. One of the most common errors in interest-rate calculations is using an interest rate that is not matched with the number of compounding periods. In equation (I.A.6.4), r/n represents the per-period interest rate and nt represents the number of compounding periods over the term of the investment. For

example, if interest is paid semi-annually then n equals 2, $r/2$ represents the semi-annual interest rate and $2t$ represents the number of semi-annual periods over the investment period.

Example I.A.6.2

Consider an investment of \$100 earning 5% per annum with interest compounded semi-annually. What is the value of the investment after three years?

The value of the investment after three years is found by applying (I.A.6.4):

$$FV = \$100 \left(1 + \frac{0.05}{2} \right)^{2 \times 3} = \$115.969$$

As the frequency of compounding increases from semi-annual to quarterly to monthly to daily to continuous compounding, it can be shown that:

$$P \left(1 + \frac{r}{n} \right)^{nt} \rightarrow P e^{rt} \tag{I.A.6.5}$$

Hence the continuous compound interest formula in equation (I.A.6.3) is the result of making the compounding periods more frequent.

I.A.6.1.3 Periodic Interest Rates versus Effective Annual Yield

Suppose you are interested in investing in a Certificate of Deposit (CD) and visit your local bank. The bank has several CDs you can purchase and lists two rates for each CD: the nominal interest rate and the effective annual yield. The nominal interest rate or annual percentage rate (APR) is the stated rate of interest for the investment. The nominal interest rate does not take into account the effect of compounding on the future value of the investment. The effective annual yield reflects the effect of compound interest on the investment for a one-year period.

Consider an investment of \$100 for a period of one year at a nominal interest rate of 5% compounded monthly. The future value of the investment is given by:

$$FV = \$100 \left(1 + \frac{0.05}{12} \right)^{12} = \$105.116$$

The effective yield, also called the annual percentage yield (APY), is calculated as:

$$EY = \frac{\text{absolute increase}}{\text{principal}} \tag{I.A.6.6}$$

A \$100 investment for one year earning 5% compounded monthly has an effective yield of:

$$EY = \frac{\$105.116 - \$100}{\$100} = 0.05116$$

The effective yield is 5.116% per annum compounded monthly, which is greater than the nominal rate of 5%. The effective yield will always be greater than the nominal rate when interest is compounded more than once a year and the effective rate is always stated as an annual rate. The effective yield will increase as the frequency of compounding increases.

An alternative formula for the effective yield when discrete compounding is used is given by:

$$EY = \left(1 + \frac{r}{n}\right)^n - 1 \quad (\text{I.A.6.7})$$

Suppose you have three guaranteed interest investment opportunities to consider. The three investments are: bank A pays 4.5% per annum compounded annually; bank B pays 4.47% per annum compounded semi-annually; and bank C pays 4.45% per annum compounded monthly. Which investment will pay the most interest? The investment opportunity with the highest effective yield will pay the most interest. Hence an investor can use the effective yield to compare investments with different interest rates and different compounding periods (see Table I.A.6.1).

Table I.A.6.1: Investment yields

	Bank A	Bank B	Bank C
Formula	$FV_A = \$1(1 + 0.045)$	$FV_B = \$1\left(1 + \frac{0.0447}{2}\right)^2$	$FV_C = \$1\left(1 + \frac{0.0445}{12}\right)^{12}$
APR	4.50%	4.47%	4.45%
Value	\$1.045	\$1.0452	\$1.04542
Effective Yield	4.5%	4.52%	4.54%

The effective yield can also be found for continuously compounded interest rates. Suppose a sum of \$100 is invested for a year at 5% per annum. At the end of one year, the investment is worth $\$100e^{0.05} = \105.127 . The effective yield can be calculated as $5.127/100 = 0.05127$.

An alternative formula for the effective yield when continuous compounding is used is given by:

$$EY = e^r - 1 \quad (\text{I.A.6.8})$$

This produces the same effective yield as equation (I.A.6.6):

$$EY = e^{0.05} - 1 \approx 0.05127.$$

I.A.6.2 Term Structure – A Definition

Our discussion of the term structure of interest rates starts with the recognition that interest rates are not assets. An investor cannot go to a market at some point in time t , buy an interest rate at a price r_t and later, at some point in time $t + \Delta t$, sell it at a price $r_{t+\Delta t}$ and realize a nominal profit or loss of $|r_t - r_{t+\Delta t}|$.

Interest rates are entities *derived* from the returns of financial assets, whose (assumed) sole purpose is to provide an investor a return for the investment of his funds. Bonds, swaps and strips² are examples of financial assets. We will carry our discussion of the term structure of interest rates by focusing on bonds. Additionally, in this section we will make the simplifying assumption of annual compounding when calculating future values.

Example I.A.6.3

Suppose you want to invest in zero-coupon bonds. You do not need to get your money back for 3 years. You have the following choices on how to proceed:

- You can buy a 3-year bond.
- You can buy a 2-year bond and, 2 years later, use its face value to buy a 1-year bond.
- You can buy a 1-year bond and, 1 year later, use its face value to buy a 2-year bond.
- You can buy a 1-year bond today and use its face value a year later to buy another 1-year bond whose face value, in turn, you use to buy a third 1-year bond, finally collecting cash at the end of 3 years.

How do you evaluate your choices? Looking at the market you find that *today* you can buy a 1-year, 2-year or 3-year bond (from issuers of similar risk), all with a face value of \$1000, at prices of \$900, \$800 and \$700, respectively.

Evaluating the simple annual returns of the bonds is straightforward. Let $B_{0,3}$ represent the present value of the 3-year bond at time 0 and $B_{3,3}$ represent the future value (face value) of the 3-year bond at maturity. In general, let $B_{n,m}$ represent the value of an m -year bond at time n .

The future value of the 3-year zero-coupon bond using annual compounding, as specified in equation (I.A.6.2), can be represented as:

$$B_{3,3} = B_{0,3} \cdot (1 + r_{0,3})^3$$

² A strip or a 'stripped bond' is a security representing a claim on either the coupon or principal of a fixed-coupon bond – see Section I.B.1.4.

where $r_{0,3}$ is the interest rate at time 0 for a 3-year investment in a zero-coupon bond. Solving for $r_{0,3}$ gives:

$$r_{0,3} = \sqrt[3]{\frac{B_{3,3}}{B_{0,3}}} - 1 = \sqrt[3]{\frac{1000}{700}} - 1 = 12.6\%$$

This is also known as the *yield to maturity* on the 3-year bond with annual compounding. (Note: the concept of bond yield is more fully covered in Chapter I.B.2.) Similarly, for the 2-year bond:

$$B_{2,2} = B_{0,2} \cdot (1 + r_{0,2})^2$$

or

$$r_{0,2} = \sqrt{\frac{B_{2,2}}{B_{0,2}}} - 1 = \sqrt{\frac{1000}{800}} - 1 = 11.8\%$$

Finally, for the 1-year bond:

$$B_{1,1} = B_{0,1} \cdot (1 + r_{0,1})$$

or

$$r_{0,1} = \frac{B_{1,1}}{B_{0,1}} - 1 = \frac{1000}{900} - 1 = 11.1\%$$

In summary, from the current market values of zero-coupon bonds of different maturities we can find the *yields to maturity* of each of the three zero-coupon bonds. Plotting these against their respective maturities (e.g. as shown in Figure I.A.6.1) gives a partial picture, albeit primitive, of the *term structure of interest rates* – also called the zero-coupon *yield curve*.

Figure I.A.6.1: The (spot) yield curve



The yield for a zero-coupon bond is also called the spot rate for the maturity of the bond. For zero-coupon bonds, the yield to maturity and the spot rate are equal. When considering coupon-paying bonds, the yield to maturity is a weighted average of the spot rates corresponding to each coupon payment. When valuing bonds, the spot rate is commonly used to calculate the present value of the cash flows because there is no concern about the reinvestment rate for the coupons received over time.

While yield curves can be constructed from prices of any set of homogeneous bonds, the yields will be influenced by factors such as credit risk, coupon rate, and frequency of coupon payments in addition to the maturity date. If the yield curve is constructed from zero-coupon government treasuries, the influence of non-maturity factors on yield is eliminated. In the USA, the government does not issue zero-coupon Treasury bonds but this limitation can be overcome by deriving the spot interest rates from coupon-paying Treasury bonds. This issue will be covered in Section I.A.6.4.

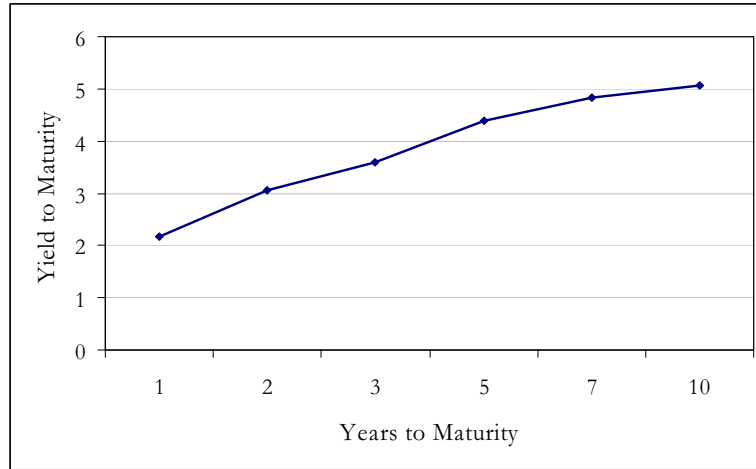
Definition: The *term structure of interest rates* or *yield curve* at time t is the relationship of bond maturity and bond yield of similar instruments. It is constructed by graphing *years* to maturity against *yield* to maturity for bonds with the same credit risk. The yield curve typically refers to the term structure of US Treasury bonds, which are considered to be default-free, although it can be constructed from non-Treasuries.

I.A.6.3 Shapes of the Yield Curve

Note that the graph of the term structure in Figure I.A.6.1 is upward sloping and is almost linear. In the real world such a graph is usually a curve and may be upward sloping, downward sloping, or partially both. Also note that in our definition we said that the term structure is defined at a specific point in time, t ; in all probability it will look different if we redraw it tomorrow or at any other time in the future. For now, especially after being told that the term structure will change, you will realize that you still cannot make a choice among your investment scenarios in Example I.A.6.3.

A yield curve represents the relationship between yield and maturity at a specific point in time. While this relationship changes over time, there are three commonly occurring yield curve shapes: normal, inverted and humped. An upward-sloping yield curve, as illustrated in Figure I.A.6.2, is the most common and is referred to as the normal yield curve. It is the result of a market expectation for higher yields as the maturity of a bond increases. Long-term bonds are generally considered to be riskier than short-term bonds. The additional risk for long-term bonds comes from the uncertainty in interest rates and yields.

Figure I.A.6.2: Normal-shaped yield curve on 31 December 2001

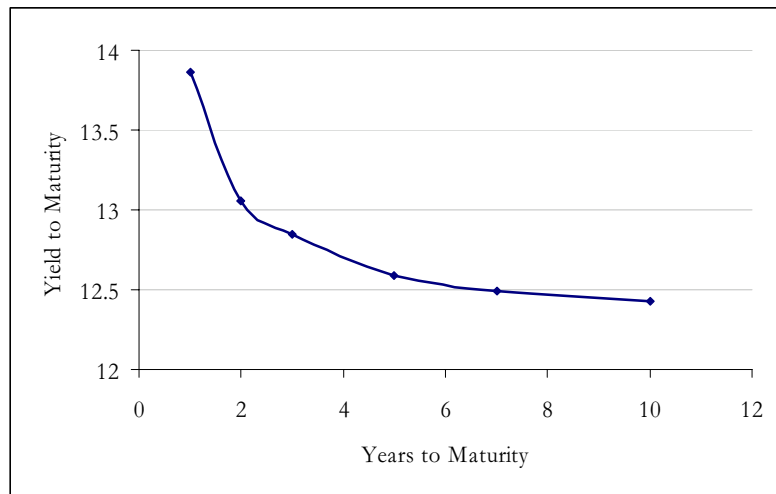


Source: TradeTools.com, US Treasuries

Inverted yield curves are unusual but not necessarily rare. Figure I.A.6.3 illustrates an inverted yield curve as observed on 31 December 1980 for US Treasuries. An inverted yield curve occurs when bonds with long maturities are expected to offer lower yields than those with short-term maturities. An inverted yield curve is generally believed to indicate an impending economic recession.

Figure I.A.6.4 illustrates a humped yield curve. The curve initially slopes up and then after a certain point in time becomes inverted. Humped yield curves are unusual and rare. Investors expect interest rates to rise initially and then decline over time. This yield curve shape generally occurs when interest rates are high compared to historical averages.

Figure I.A.6.3: Inverted yield curve on 31 December 1980



Source: TradeTools.com, US Treasuries

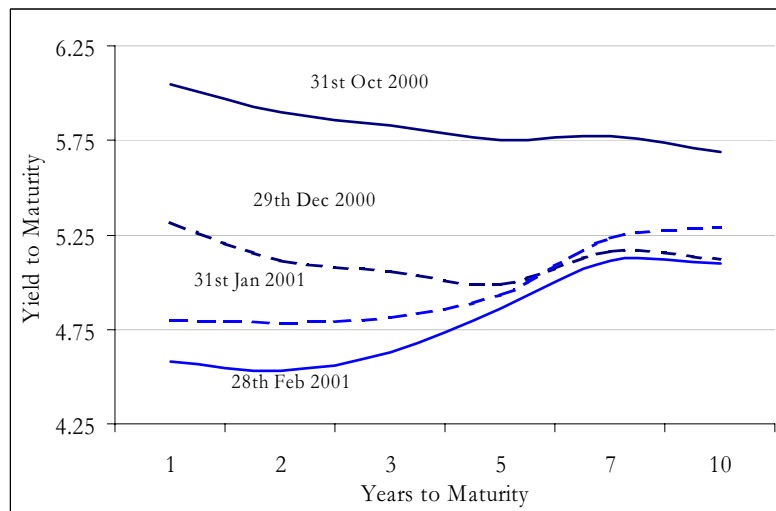
Figure I.A.6.4: Humped yield curve on 1 May 1989



Source: TradeTools.com, US Treasuries

In Figure I.A.6.5 the transition from an inverted yield curve on 31 October 2000 to a normal upward-sloping yield curve on 28 February 2001 is illustrated. As investors change their expectations of the relationship between long and short-term interest rates, the yield curve can take on strange shapes.

Figure I.A.6.5: Transitional yield curves



Source: TradeTools.com, US Treasuries

I.A.6.4 Spot and Forward Rates

Most consumers are familiar with financial transactions involving market spot rates. When financing a car or a home, loan payments are generally determined using the prevailing interest rates. But there are situations when a consumer is interested in what will happen to interest rates in the future. Suppose you purchase a house and have two financing options: a 30-year fixed rate at 6.4% or a 7-year adjustable-rate mortgage at 5.5%. If you select the 7-year adjustable rate mortgage, you are concerned with what happens to interest rates in the future as the interest rate applied to the outstanding principal changes in line with variations in a benchmark interest rate such as LIBOR.

Similarly, in many commercial transactions borrowers and lenders agree now to make a loan in the future. A forward rate contract on interest rates is known as a forward rate agreement and is discussed in Section I.A.7.3.

Consider again the four zero-coupon investment strategies introduced in Section I.A.6.2. Suppose you knew with 100% certainty that if you were to buy a 2-year bond one year from now it would cost you \$780, while a 1-year bond would cost you \$740. Also, if you were to buy a 1-year bond two years from now it would cost you \$840. Within minutes you have all your calculations done. You can see how you got the yields shown in Table I.A.6.2. For example, the yield of a 1-year bond bought one year from now is given by:

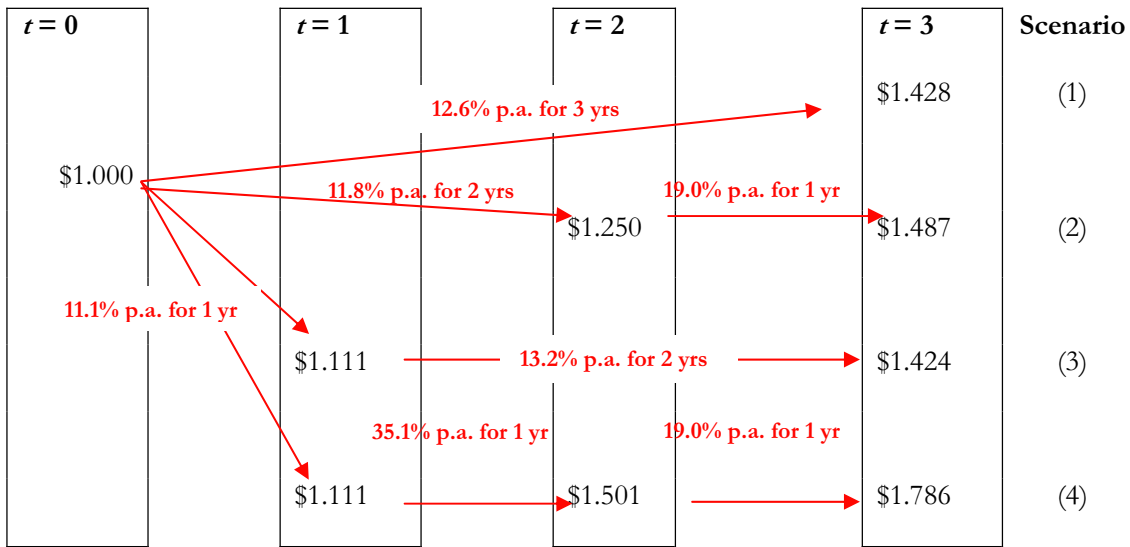
$$\frac{B_{1,2}}{B_{1,1}} - 1 = \frac{1000}{740} - 1 = 35.1\%$$

Table I.A.6.2: Bond yields versus maturity

	Matures at $t =$		
Bought at $t =$	1	2	3
0	11.1%	11.8%	12.6%
1		35.1%	13.2%
2			19.0%

So, if you were to invest \$1 today, and possibly reinvest your payoff, depending on your choice, at the end of 3 years your total payoff would be any of the four shown in Figure I.A.6.6.

Figure I.A.6.6: Returns for four alternative scenarios of investing in bonds for 3 years



In a world with fixed or certain future interest rates, your investment choice would be easy: scenario (4) in Figure I.A.6.6 would give you the highest payoff for your investment of \$1,786. This equates to an average annual yield of $(\sqrt[3]{1.786} - 1) = 21.3\%$ versus 12.6%, 14.1% and 12.5% for scenarios (1)–(3), respectively.

There are two types of interest rates illustrated in the example above.

- The *spot rate* is the prevailing interest rate on a zero-coupon bond. As in Example I.A.6.3, $r_{0,t}$ denotes a spot rate that is observed today and continues until time t . In Figure I.A.6.6, the spot rate for a 1-year bond at time 0 is 11.1% and the spot rate for a 3-year bond at time 0 is 12.6%. When valuing bonds, the spot rate is commonly used to calculate the present value of the cash flows because there is no concern about the reinvestment rate for the coupons received over time. We also use the notation $r_{x,t}$ for $x > 0$, to denote the spot rate that is observed at time $x > 0$ and prevails until time t .
- A *short rate* is a *one-period* interest rate that prevails at some current or future time t . The current one-period spot interest rate is $r_{0,1}$ and, more generally, the short rate at some future time $t - 1$ that prevails during the t th year is $r_{t-1,t}$. Future short rates are unknown, but in a world with no uncertainty a future short rate $r_{t-1,t}$ that is ‘solved’ or ‘implied’ by the current spot rates is also called the *forward rate*, $f_{t-1,t}$.³ Note that forward rates are marginal rates, whereas a yield is an average rate.

³ Section I.A.6.5 will examine possible relationships between forward rates and future short rates when future short rates are uncertain.

Table I.A.6.3: Forward rates versus maturity

Forward rates (%)	Maturity (years)
2	1
4	2
5	3

Let us explore the relationship between the spot rate and short rates, assuming that future short rates are known. Suppose that in the interest-rate market today, the forward rates in Table I.A.6.3 are available. The current spot rates are implied by the known forward rates. One dollar invested at the 2-year spot rate will be worth $\$1(1+r_{0,2})^2$ at the end of the 2-year term. Similarly, if you invest \$1 at the 1-year spot rate, $r_{0,1}$, and roll over the proceeds at the prevailing forward rate, $r_{1,2}$, at the end of the 2-year period, your investment is worth $\$1(1+r_{0,1})(1+r_{1,2})$. These two investment alternatives need to produce the same payoff, under the assumption of known forward rates, otherwise a risk-free profit can be made. Hence

$$\$1(1+r_{0,2})^2 = \$1(1+r_{0,1})(1+f_{1,2})$$

or

$$\$1(1+r_{0,2})^2 = \$1(1+0.02)(1+0.04).$$

Solving the above equation gives the current spot rate a value of 2.995% derived from the known forward rates.

Example I.A.6.4

Find the 3-year spot rate given the forward rates in Table I.A.6.3.

Investing \$1 at the prevailing spot rate for a 3-year period has a value of $\$1(1+r_{0,3})^3$. Similarly, investing the \$1 for 1-year periods with the proceeds rolled over at the prevailing short rate will be worth $\$1(1+r_{0,1})(1+r_{1,2})(1+r_{2,3})$ at the end of the 3-year period. Both investments must produce the same payoff, therefore

$$\$1(1+r_{0,3})^3 = \$1(1+r_{0,1})(1+r_{1,2})(1+r_{2,3})$$

or

$$\$1(1+r_{0,3})^3 = \$1(1+0.02)(1+0.04)(1+0.05).$$

Solving this gives a spot rate of 3.659%.

There are two important observations from the above examples. First, the spot rate is a geometric average of the short rates. Second, the yield curve is a graph of the spot rates for bonds with identical risk and different maturities.

Table I.A.6.4: Spot rates versus maturity

Spot rates (%)	Maturity (years)
2	1
4	2
6	3

Now consider turning the situation around – we know the spot rates and want to find the forward rates. Assume the spot rates in Table I.A.6.4. Investing \$1 for a 2-year term at 4% must be equivalent to investing for a year at the 1-year short rate, 2%, and reinvesting the proceeds for a second year at the short rate, $r_{1,2}$.

$$\$1(1 + r_{0,2})^2 = \$1(1 + r_{0,1})(1 + r_{1,2})$$

or

$$\$1(1 + 0.04)^2 = \$1(1 + 0.02)(1 + r_{1,2}).$$

Solving for $r_{1,2}$ gives a 1-year short rate starting in year 2 of 6.04%.

Example I.A.6.5

Find the 1-year short rate starting in year 2 using the data in Table I.A.6.4.

Investing \$1 for 3 years at the prevailing spot rate must result in the same payoff as investing at the certain short rates for 1-year terms:

$$\$1(1 + r_{0,3})^3 = \$1(1 + r_{0,1})(1 + r_{1,2})(1 + r_{2,3})$$

or

$$\$1(1 + 0.04)^3 = \$1(1 + 0.02)(1 + 0.0604)(1 + r_{2,3}).$$

Solving for $r_{2,3}$ shows the short rate at the end of year 2 is approximately 4%.

We need to clarify that the forward rate for year t is a rate *implied* from a spot yield curve. While it is not always an observable rate of interest, people use the forward rate to denote what the ‘prevailing’ interest rate for a certain year ‘will be’. Suppose the short rates for years 1, 2 and 3 are known and denoted as $r_{0,1}$, $r_{1,2}$ and $r_{2,3}$, respectively; then the following relationship will hold:

$$\begin{aligned}
 (1 + r_{0,2})^2 &= (1 + r_{0,1})(1 + r_{1,2}) \\
 (1 + r_{0,3})^3 &= (1 + r_{0,1})(1 + r_{1,2})(1 + r_{2,3}) \\
 &\vdots \\
 (1 + r_{0,n})^n &= (1 + r_{0,1})(1 + r_{1,2}) \cdots (1 + r_{n-1,n}).
 \end{aligned}
 \tag{I.A.6.9}$$

If we do not know a certain short rate, we can ‘discover’ it by solving the above equations. For example, if we knew the yields for a 1-year bond bought at time $t = 0$ and a 2-year bond bought at time $t = 0$, we can use:

$$[1 + r_{0,1}] \cdot [1 + r_{1,2}] = [1 + r_{0,2}]^2$$

to get the short rate at the end of year 1 that will prevail during year 2:

$$r_{1,2} = \frac{[1 + r_{0,2}]^2}{1 + r_{0,1}} - 1.$$

In general, the short rate for year t is given by the recursive formula:

$$r_{t-1,t} = \frac{[1 + r_{0,t}]^t}{[1 + r_{0,t-1}]^{t-1}} - 1. \tag{I.A.6.10}$$

In the real world, which is dynamic, the yields of bonds of any maturity change over time and as such the forward rate also changes. Suppose you wanted today to calculate the forward rate that will prevail during the third year down the road. You find that the yield of a bond bought today and maturing in 3 years is 12.3%, while the yield of a bond bought today and maturing in 2 years is 11.6%. Your calculation gives you:

$$\frac{[1 + 0.123]^3}{[1 + 0.116]^2} - 1 = 0.137 \text{ or } 13.7\%.$$

Three months later, the prices of these same bonds have changed and their yields are now found to be 12.7% and 11.2%. Repeating your calculation gives you:

$$\frac{[1 + 0.127]^3}{[1 + 0.112]^2} - 1 = 0.158 \text{ or } 15.8\%.$$

So a change in the bond yields of the order of only 3-4% results in more than a 15% increase in the short rate. You may have a hint now that our implied values for future short rates are valid only at the moment we make the calculation. For this reason, in Section I.A.6.5 we will model future short rates as random variables.

Example I.A.6.6: Estimating spot rates from coupon paying US Treasury bonds

Suppose that you see in the *Wall Street Journal* a list of yields to maturity for US Treasury bonds as in Table I.A.6.5.

Table I.A.6.5: US Treasuries yields to maturity

Maturity (years)	Price	Yield to maturity
1	100	5.00%
2	100	5.35%
3	100	5.75%

The spot rates can be determined from the yield-to-maturity data of coupon-paying Treasury bonds by using an iterative process commonly referred to as *bootstrapping*. The price of a bond is simply the present value of the cash flows. The 2-year spot rate $r_{0,2}$ can be found by solving the following equation:

$$P_2 = \frac{\text{coupon}_1}{(1+r_{0,1})} + \frac{\text{face} + \text{coupon}_2}{(1+r_{0,2})^2} = \frac{5.35}{1+0.05} + \frac{100+5.35}{(1+r_{0,2})^2}$$

Solving for $r_{0,2}$ gives:

$$r_{0,2} = \left(\frac{105.35}{100 - 5.35/1.05} \right)^{0.5} - 1 \approx 0.053592$$

or approximately 5.3592%, which is 0.0092% higher than the yield. Similarly, the 3-year spot rate can be found by solving the following equation for $r_{0,3}$:

$$100 = \frac{5.75}{1+0.05} + \frac{5.75}{(1+0.053592)^2} + \frac{105.75}{(1+r_{0,3})^3}$$

Solving the equation for $r_{0,3}$ gives approximately 5.7804%, which is 0.0304% higher than the yield.

Note that all spot rates are higher than the yield to maturity and that the difference between the yield to maturity and the spot rate increases as the maturity of the bond increases.

I.A.6.5 Term Structure Theories

Suppose you are the chief financial officer of ABC Corporation. The company is planning to finance a major plant expansion project by issuing floating-rate notes (see Section I.B.1.4). As the CFO you are concerned about what happens to interest rates and you are estimating the impact of a change in interest rates on the feasibility of the project. You determine that if interest rates increase by 2%, the interest expense for the project climbs significantly and the project is no longer profitable. You are faced with the question of whether or not to hedge your interest-rate exposure. There are two optimal scenarios: you decide to hedge and interest rates go up; or you decide not to hedge and interest rates go down. Whether you hedge or not will be determined by your *expectations* concerning future changes in short interest rates. If you expect interest rates or inflation to go up, relative to the rates already built into the forward yield curve, you will hedge your interest-rate exposure. There are primarily three theories used to explain the shape of the yield curve, that is, what will happen to future interest rates. Two are based on the market's expectation of future spot rates.

I.A.6.5.1 Pure or Unbiased Expectations

The pure expectations hypothesis of the shape of the yield curve is based on the belief that the forward rate for a particular time period is the best predictor of the expected short rate. That is,

$$f_{t-1,t} = E(r_{t-1,t}) \quad (\text{I.A.6.4})$$

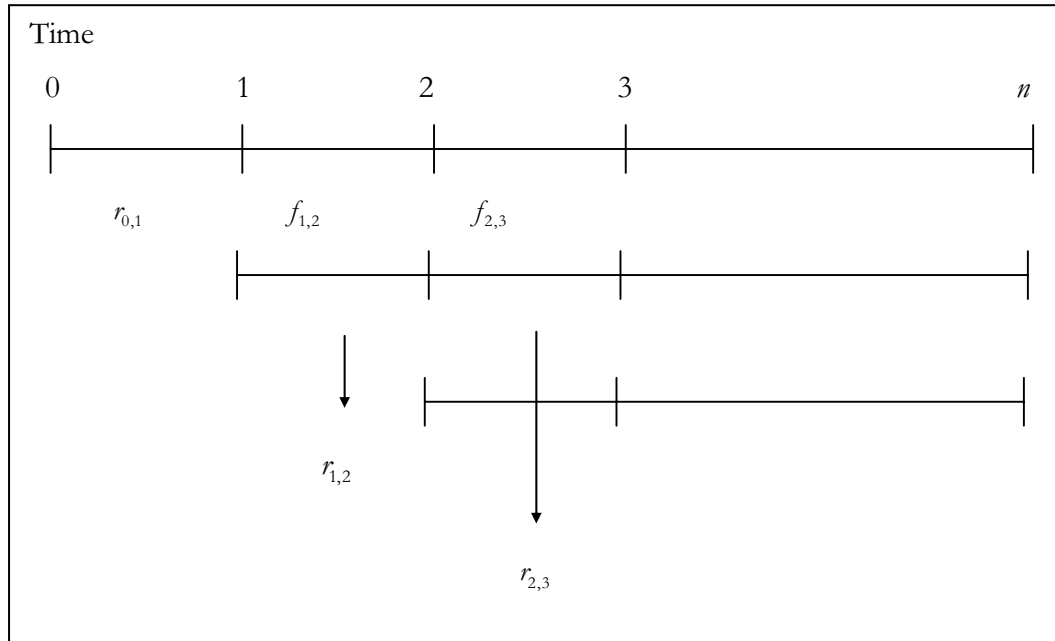
Figure I.A.6.7 illustrates the relationship between forward rates at time 0 and the unknown future spot rates. At time 0, the forward rate, $f_{1,2}$, is considered to be the *best predictor* of the future (unknown) spot rate for the same time period.

The pure expectations hypothesis can be used to explain any yield curve shape. If investors expect next period's short rate to be higher than this period's short rate, the yield curve will be upward sloping. Conversely, if investors expect next period's short rate to be lower than this period's short rate, the yield curve will be downward sloping. A flat yield curve will occur when investors expect future interest rates will equal the current spot rate.

Expectations of future interest rates are highly dependent on investors' expectations of inflation. When inflation is expected to increase, longer-term rates will be higher than short-term rates. Hence, according to the pure expectations hypothesis, the shape of the yield curve is dependent on investors' expectations of future interest rates and inflation as well as other macroeconomic factors.

For the CFO mentioned above, analysis of hedging strategies is relatively straightforward under the pure expectations hypothesis. The forward yield curve represents the market’s perceptions of future rates. If the CFO believes that interest rates will increase beyond the levels represented in the forward curve, then he will hedge.⁴

Figure I.A.6.7: Relationship between forward and future spot interest rates



I.A.6.5.2 Liquidity Preference

The liquidity preference hypothesis of the shape of the yield curve is an extension of the pure expectations hypothesis by factoring in investor risk aversion and uncertainty:

$$f_{t-1,t} = E(r_{t-1,t}) + p_t \quad (\text{I.A.6.11})$$

In equation (I.A.6.11), investor risk aversion and uncertainty are contained in the *risk premium* term, p_t . Note that the size of the risk premium can vary across time, complicating the analysis further. Research studies have found evidence of positive risk premiums, suggesting that investors require a higher return in order to commit their money for longer time periods.

If the liquidity preference hypothesis holds, then an upward-sloping forward curve does not necessarily mean that market participants expect interest rates to increase. A flat yield curve might actually indicate an expectation of constant future interest rates, after taking account of the liquidity premium.

⁴ Of course, this assumes that the CFO has reason to believe that he can predict interest rates better than the rest of the market!

I.A.6.5.3 Market Segmentation

The market segmentation theory states that investors in long-term fixed income securities are different from investors in short-term fixed income securities. Investors decide on the term of their investments based on their future need for liquidity and their risk preferences. According to the market segmentations theory, the shape of the yield curve is determined by the supply-and-demand dynamics of long-term versus short-term fixed income investors.

I.A.6.6 Summary

Interest-rate calculations are a fundamental necessity of everyday life. When a consumer purchases a car or home, or a credit card user decides not to pay off the balance at the end of the month, compounding frequency on the outstanding debt becomes a very important issue. Compounding interest on a daily, weekly or monthly basis will result in very different total amounts paid over the life of the loan. The more frequent the compounding period, the more interest paid over the life of the loan, which is profitable for the lender and costly for the borrower.

Similarly, the yield curve is a very important tool for financial engineers. Many asset pricing methodologies are based on discounted cash-flow analysis, which relies on a discount factor and hence an interest rate. Bond and derivative pricing are just two examples that rely on a specific yield curve – the zero-rate yield curve – for accurate pricing. Whereas a yield curve can be constructed from any set of homogeneous bonds, the zero-rate curve is constructed by bootstrapping the zero rates from a set of coupon-paying government treasury bonds. The zero rates are clean rates that only reflect the time value of money, whereas other rates are contaminated with coupon payments. The zero rates are needed to derive a fair price for an asset.

I.A.7 Valuing Forward Contracts

Don Chance¹

This chapter covers the principles for valuing and pricing forward contracts. It begins by highlighting the distinction between pricing and valuation of forward contracts in Section I.A.7.1. The material is then split into two further parts according to the nature of the underlying. Pricing and valuation principles for forward contracts where the underlying is an asset (such as a stock, bond, commodity or currency) are covered in Section I.A.7.2. Similar pricing and valuation principles where the underlying is an interest rate are the topic of Section I.A.7.3.

Within *The PRM Handbook* there are two other closely related sections covering forward/futures contracts and markets. Futures and forwards are also discussed in Chapter I.B.3, which focuses on applications of these important financial instruments. Finally, a discussion of the characteristics and operation of the market for futures may be found in Chapter I.C.6.

I.A.7.1 The Difference between Pricing and Valuation for Forward Contracts

In markets for assets such as stocks, bonds, currencies, commodities and options, the *price* of an asset is defined as the amount of money that a buyer would pay and a seller would receive to enter into a transaction in the asset. The *value* of the asset is the amount of money that a buyer or seller believes the asset is worth. Value is found by either an assessment of the asset's expected cash flows and risk or by comparing the asset's cash flows to those of an instrument with a known price and identical cash flows. An asset with a price less (more) than its value is underpriced (overpriced). In a reasonably well-functioning market, overpriced assets are quickly sold and underpriced assets are quickly purchased, resulting in a convergence of price to value. In a truly efficient market, this convergence occurs so rapidly that no single participant can consistently exploit any discrepancies.

This standard notion of the relationship between price and value, however, fails when we are dealing with forward contracts, futures contracts and swaps. These instruments are commitments for one party to purchase and the counterparty to sell an asset for an amount of money agreed upon at the start.² These contracts expire at a specific date. The amount of money that will be

¹ Don M. Chance, Ph.D., CFA. William H. Wright, Jr. Chair of Financial Services, Department of Finance, 2163 CEBA, Louisiana State University

² While we refer to these instruments as commitments to purchase an *asset*, there is no reason why the underlying cannot be another forward, future or swap, which is not an asset at the start.

exchanged at the expiration date is referred to as the *price*. This money is not, however, exchanged at the start of the contract. For that reason, the notion of a *price* associated with forwards, futures and swaps is an entirely different concept from the price of an asset, which is indeed exchanged at the start.

The value of a forward, future or swap is the amount of money that the buyer or seller believes the contract is worth and represents the amount of money that the buyer would be willing to pay and the seller would be willing to receive to engage in the contract. That value is obtained by comparing the cash flows from the contract with those of a transaction with a known price that produces identical cash flows. The value obtained will then imply a fair price at which a counterparty should be willing to engage in the contract.

As a result of this different notion of price, the price of a forward, future or swap is typically of an entirely different order of magnitude than the value of the contract. It will be the case, however, that, if the price at which a counterparty offers to engage in the contract differs from the price implied by the value, then the contract is mispriced and an arbitrage opportunity is available.

Valuation is particularly important given that accounting practice either requires or is likely to require (depending on the country) that derivatives values be included in financial statements. Our focus here is on forward contracts, but, as noted, similar principles apply to futures and swap contracts. We divide the material into two sections, one in which the underlying is an asset, such as a stock, bond, currency or commodity, and another in which the underlying is an interest rate.

I.A.7.2 Principles of Pricing and Valuation for Forward Contracts on Assets

We begin with notation. Consider a forward contract established today at time 0. The contract expires at time T , which can be interpreted as the number of years, or alternatively $T = (\text{days to expiration}/365)$. The forward price, which is the price agreed upon today, is denoted as $F(0,T)$. The price of the underlying asset, known as the spot price, is denoted as S_0 today and S_T at the expiration date. We shall also be interested in an intermediate time point denoted as time t , where $t = (\text{days until that point in time})/365$. The spot price at t is S_t . We assume that the default-free interest rate is a constant r , compounded annually.

The forward price is the aforementioned $F(0,T)$. If a new forward contract is created at t to expire at T , its price would be denoted as $F(t,T)$. A new forward contract created at T to expire at T

would have a price of $F(T,T)$. Such a contract would be equivalent to a spot transaction, however, so $F(T,T) = S_T$.

The value of a forward contract created at time 0 is $V_0(0,T)$. The value of that same contract at t is $V_t(0,T)$, and its value at T is $V_T(0,T)$. Note that a subscript indicates value at a point in time, while the arguments in parentheses indicate that the contract was created at time 0 and expires at time T .

The two parties are referred to as the buyer or the *long* and the seller or the *short*. We ignore any credit considerations by assuming that each party is default-free or fully insured at no cost. All cases assume a forward contract on one unit of the underlying. If the contract covers more than one unit of the underlying (as is always the case), the value of the contract is multiplied by the number of units, although the price is typically stated on a per-unit basis. In addition, all results are obtained from the point of view of the party holding the long position. The forward price is the same for the long and the short. The value for the short is minus one times the value for the long.

I.A.7.2.1 The Value at Time 0 of a Forward Contract

Because no money changes hands at the time the forward contract is created (time 0), the value at time zero has to be zero. Hence, we obtain our first result:

$$V_0(0,T) = 0 \tag{I.A.7.1}$$

If a party perceives that the contract has a value other than zero, it would engage in the contract and capture a gain equal to the non-zero value. To obtain a value of zero, however, the forward contract price must be a specific amount. We determine the appropriate forward contract price in Section I.A.7.2.5.

I.A.7.2.2 The Value at Expiration of a Forward Contract on an Asset

At expiration, the spot price is S_T . The long holds a contract requiring purchase of the asset worth S_T at a price of $F(0,T)$. The value to the long is, therefore,

$$V_T(0,T) = S_T - F(0,T) \tag{I.A.7.2}$$

For example, if the long were obligated to purchase an asset for \$100 and the asset is selling for \$105, the contract is worth \$5 to the long. The long should be able to sell the contract for \$5, pledge it as \$5 of collateral or engage in the transaction adding \$5 of value to its net worth.

I.A.7.2.3 The Value Prior to Expiration of a Forward Contract on an Asset

As we have seen, it is a simple matter to determine the value of a forward contract at the time it is initiated and at expiration. It is slightly more difficult to determine its value during the life of the contract.

Suppose we are now at time t , part of the way through the life of the contract. We can determine the value of the forward contract at t if we can find an alternative transaction that is guaranteed to produce the same outcome at time T as the forward contract. Suppose at time t , we purchase the asset and borrow $F(0,T)$ using a discount loan. The cost of this transaction is $S_t - F(0,T)(1 + r)^{-(T-t)}$, which can also be viewed as the value of the transaction of buying the asset and borrowing at T . At time T , we will be holding an asset worth S_T and will owe the amount $F(0,T)$. This is the same as the value of the forward contract at time T . Hence, the value of the forward contract at time t is the same as the value of the transaction of buying the asset and borrowing $F(0,T)$. Thus,

$$V_t(0,T) = S_t - F(0,T)(1 + r)^{-(T-t)} \quad (\text{I.A.7.3})$$

Note that equation (I.A.7.2) is the same as I.A.7.3 when $t = T$. That is, when we let t go all the way to T , our valuation formula (I.A.7.3) converges to the simple formula (I.A.7.2). As we shall see, our formula for the value at time 0, equation (I.A.7.1), is a special case of equation (I.A.7.3).

I.A.7.2.4 The Value of a Forward Contract on an Asset when there are Cash Flows on the Asset during the Life of the Contract

In the example above, the underlying asset generated no cash inflows or outflows, nor did it incur any explicit holding costs. Many assets generate cash inflows or outflows. Most stocks pay dividends, most bonds pay interest, and all currencies pay interest, however little it may be at times. While these assets typically do not generate significant costs to holding them, most commodities do incur such costs. In addition, some assets, particularly certain commodities, generate certain non-monetary benefits.

The explicit and implicit costs associated with holding an asset will affect the arguments presented above. For example, let us consider an asset such as a stock or bond that makes cash payments (dividends for stocks, coupons for bonds). These cash payments have a present value at time t of $C(t,T)$. Interpret this variable as the present value at time t of the cash payments over the period from t to T . Let us return to our argument that buying the asset and borrowing $F(0,T)$ at t replicates the value of the forward contract at T . Now it no longer does. The holder of the asset will not only have the asset worth S_T but will also have $C(t,T)(1 + r)^{T-t}$, which can be interpreted as the value at T of the cash payments collected and reinvested in the risk-free asset over the period from t to T .

To recover the correct valuation formula, let us alter our strategy slightly. At time t , we borrow $C(t,T)$ as well as $F(0,T)(1+r)^{-(T-t)}$. So the value of this strategy at t is

$$S_t - C(t,T) - F(0,T)(1+r)^{-(T-t)}$$

At time T , we shall have S_T (the asset price) + $C(t,T)(1+r)^{T-t}$ (the accumulated and reinvested cash payments) – $C(t,T)(1+r)^{T-t}$ (the repayment of one loan) – $F(0,T)$ (the repayment of the other loan), which equals $S_T - F(0,T)$ and now equates to the value of the forward contract. Thus, *when the underlying is an asset that makes cash payments*,

$$V_t(0,T) = S_t - C(t,T) - F(0,T)(1+r)^{-(T-t)} \quad (\text{I.A.7.4})$$

The formula for the value at time 0, equation (I.A.7.1), is not altered by the presence of these cash payments. The value of the contract is zero at the start, because no money changes hands, whether these cash payments are made or not. At expiration, the value formula, equation (I.A.7.2), is the same because there are no cash payments remaining.

If the underlying is a currency, we typically capture the information in the cash payments in the form of a foreign interest rate, which we denote as r_f . Now note that, at time T , the party buying the currency and borrowing at time t will have not one unit of the currency but $(1+r_f)^{T-t}$ units of the currency. Thus, the value of this strategy at T will be $S_T(1+r_f)^{T-t} - F(0,T)$. Naturally, this value exceeds that of the forward contract at T . If we alter the strategy at t so that instead of buying one unit of the asset, we buy $(1+r_f)^{-(T-t)}$ units, we shall have

$$(1+r_f)^{-(T-t)}S_T(1+r_f)^{T-t} - F(0,T) = S_T - F(0,T),$$

which now matches the value of the forward contract at expiration. Thus, the value of a currency forward contract at t is as follows: *When the underlying is a currency that pays interest at the rate r_f* ,

$$V_t(0,T) = S_t(1+r_f)^{-(T-t)} - F(0,T)(1+r)^{-(T-t)} \quad (\text{I.A.7.5})$$

As in the case of stocks and bonds, the values at times 0 and T are not affected by interest on the currency.

Now suppose the underlying is a commodity that incurs storage costs. Let us define these costs in terms of a known amount that has a present value at t of $\gamma(t,T)$ (the Greek letter gamma). At time T , these costs will have a value of $\gamma(t,T)(1+r)^{T-t}$. This value, $\gamma(t,T)(1+r)^{T-t}$, can be viewed as the amount of money expended for storage (including the interest forgone) over the period from t to T . The strategy we described at time t will now produce a value at T of $S_T - \gamma(t,T)(1+r)^{T-t} - F(0,T)$, which differs from the value of the forward contract at T . Now we must alter the strategy at t to add an investment in the risk-free bond in the amount of $\gamma(t,T)$. Then at T , the

value of the strategy is S_T (from the asset) $- \gamma(t,T)(1+r)^{T-t}$ (the accumulated storage costs on the asset) $+ \gamma(t,T)(1+r)^{T-t}$ (the payoff of the risk-free bond) $- F(0,T)$ (the payment of the loan) $= S_T - F(0,T)$, which now matches the value of the forward contract. So this strategy replicates the forward contract on the commodity. Thus, its value is the same as the value of the forward contract at t : *When the underlying is a commodity that incurs storage costs,*

$$V_t(0,T) = S_t + \gamma(t,T) - F(0,T)(1+r)^{-(T-t)} \quad (\text{I.A.7.6})$$

Note how equation (I.A.7.6) can be linked to equation (I.A.7.4), the valuation formula when there are cash flows. If there are positive cash flows, we typically refer to them as dividends or coupons and, accordingly, we obtain equation (I.A.7.4). Negative cash flows are referred to as holding costs and lead to equation (I.A.7.6). But there is technically no reason why a single formula would not apply in both cases. The sign of the cash flows could be positive (for dividends or coupons) or negative (for holding costs). But for convenience and distinction, we shall use separate formulae.

Finally, we must note that some commodities pay an implicit return known as a convenience yield. The convenience yield is a concept not very well understood but is thought to represent a form of non-pecuniary return often occurring in markets with shortages. Loosely speaking, the convenience yield is a form of a premium earned by holders of assets in short supply. It is a simple matter to incorporate a convenience yield into equation (I.A.7.6). We merely redefine the storage cost $\gamma(t,T)$ to be the storage cost net of convenience yield. In some cases, the convenience yield could exceed the storage cost, giving $\gamma(t,T)$ a negative sign and making equation (I.A.7.6) more like equation (I.A.7.4).

Again, the values of the forward contract at times 0 and T are unaffected by the presence of storage costs and a convenience yield.

I.A.7.2.5 Establishing the Price of a Forward Contract on an Asset

Having determined the value of a forward contract at various times during its life, we can easily determine the price, $F(0,T)$, established when the contract is initiated. Returning to the simple case of no cash flows or storage costs, we know that the value of a forward contract at time t is given by equation (I.A.7.3). We know that at time 0, however, no money changes hands and the value must be zero, as indicated in equation (I.A.7.1). Thus, these two equations must be equal. Setting t to 0, equation (I.A.7.3) to 0 and solving for $F(0,T)$, we obtain

$$F(0,T) = S_0(1+r)^T \quad (\text{I.A.7.7})$$

The forward price for the contract is thus set at the compound future value of the spot price.

A somewhat more typical approach to solving for $F(0,T)$ is simply to propose that an individual purchase the asset at time 0 and sell a forward contract. This transaction will guarantee that at T the individual will receive $F(0,T)$ for the asset. This transaction is, therefore, risk-free so the future payoff of the asset must provide for a return equal to the risk-free rate. Hence, $F(0,T)$ must equal $S_0(1 + r)^T$.

If a party would enter into a long position in a forward contract at a price in excess of $S_0(1 + r)^T$, the counterparty going short will earn more than the risk-free rate. The short could borrow the funds at the risk-free rate, buy the asset and earn an infinite profit at no risk. The demand for short positions in forward contracts would be infinite, driving the market price down until it equals $S_0(1 + r)^T$. Similarly, if a participant would sell a contract at a price below $S_0(1 + r)^T$, other parties would buy the contract and sell short the asset, creating a synthetic risk-free loan that would cost less than the risk-free rate. The funds could be invested to earn the risk-free rate, leading to an infinite rate of return, infinite demand for long positions in the forward contract and an ensuing increase in the market price of the forward contract until it equalled $S_0(1 + r)^T$.

A similar line of reasoning can be applied to the cases in which the underlying pays dividends, coupons or foreign interest, or incurs storage costs. The resulting forward prices are, *when the underlying pays dividends or coupons,*

$$F(0,T) = (S_0 - C(0,T))(1 + r)^T \quad (\text{I.A.7.8})$$

When the underlying is a currency that pays interest at the rate r_f ,

$$F(0,T) = S_0(1 + r)^{-T} (1 + r_f)^T \quad (\text{I.A.7.9})$$

When the underlying is a commodity that incurs storage costs,

$$F(0,T) = (S_0 + \gamma(0,T))(1 + r)^T \quad (\text{I.A.7.10})$$

In general, these formulae can be characterised as the spot price, net of any benefits or plus any costs, compounded at the risk-free rate over the life of the contract.

I.A.7.2.6 Pricing and Valuation when the Cash Flows or Holding Costs are Continuous

In some circumstances, it is best to work with continuous rates.³ The continuously compounded interest rate, which we denote as r^c , is the rate at which interest grows when compounded an infinite number of times during a finite period. This rate is found as the simple logarithmic transformation of the discrete rate r , that is

³ See the discussion of discrete versus continuous compounding methods in Chapter I.B.1.

$$r^c = \ln(1 + r)$$

When interest is compounded continuously, our preceding formulae are slightly adjusted. For example, the formulae for the forward price established at time 0 and value at time t when the asset has no cash flows or costs are

$$\begin{aligned} F(0, T) &= S_0 e^{r^c T} \\ V_t(0, T) &= S_t - F(0, T) e^{-r^c (T-t)} \end{aligned}$$

When the underlying generates cash flows, it is common to use $C(0, T)$ as the present value of the cash flows when these cash flows are in the form of bond coupons. In that case our formulae would be

$$\begin{aligned} F(0, T) &= (S_0 - C(0, T)) e^{r^c T} \\ V_t(0, T) &= S_t - C(t, T) - F(0, T) e^{-r^c (T-t)} \end{aligned}$$

When the underlying is a stock, we could use $C(t, T)$ as above, but it is more typical to incorporate the continuously compounded dividend yield, which we denote as δ^c . In that case, we must remove the value of the dividends from the stock price, an operation performed by discounting the stock price at the dividend yield rate over the contract life. Thus, our price and valuation formulae become

$$\begin{aligned} F(0, T) &= S_0 e^{-\delta^c T} e^{r^c T} = S_0 e^{(r^c - \delta^c) T} \\ V_t(0, T) &= S_t e^{-\delta^c (T-t)} - F(0, T) e^{-r^c (T-t)} \end{aligned} \tag{I.A.7.11}$$

When the underlying is a currency, we simply use the continuously compounded foreign rate, which we denote as r_f^c . Then our formulae become

$$\begin{aligned} F(0, T) &= S_0 e^{-r_f^c T} e^{r^c T} = S_0 e^{(r^c - r_f^c) T} \\ V_t(0, T) &= S_t e^{-r_f^c (T-t)} - F(0, T) e^{-r^c (T-t)} \end{aligned}$$

Note the similarity between the formulae for a forward contract on a currency with foreign interest rate r_f^c and those for a forward contract on a stock with dividend yield δ^c . The foreign interest rate and dividend yield play the same role. This result should make sense. The foreign interest rate and the dividend yield are the rates at which their respective underlyings generate positive cash flows.

When the underlying is a commodity with storage costs, we can continue to use the measure of storage costs $\gamma(0, T)$ at time 0 and $\gamma(t, T)$ at time t . In that case, the formulae become

$$F(0, T) = (S_0 + \gamma(0, T))e^{r^f T}$$

$$V_t(0, T) = S_t + \gamma(t, T) - F(0, T)e^{-r^f(T-t)}$$

Alternatively, we might prefer a continuously compounded measure of the rate at which storage costs accrue. Denoting this parameter as λ , our formulae would be

$$F(0, T) = S_0 e^{\lambda T} e^{r^f T} = S_0 e^{(r^f + \lambda)T}$$

$$V_t(0, T) = S_t e^{\lambda(T-t)} - F(0, T) e^{-r^f(T-t)}$$

In this case, note how the storage cost rate λ is added to the risk-free rate to determine the forward price. The intuition is simple: interest is a form of storage cost, representing the opportunity cost of funds tied up in the asset. The rates r and λ constitute the total cost forgone in holding the asset.

I.A.7.2.7 Numerical Examples

In this section, we present some numerical examples. All cases involve discrete compounding. Conversion to the continuous case can be done easily by adjusting the rates and measuring the costs and cash flows with continuous compounding and discounting.

We take as the standard case an asset priced at €100. The discrete risk-free rate is 4%. We are interested in a forward contract expiring in three years. Thus, $S_0 = 100$, $r = 0.04$, and $T = 3$. We shall find the forward price, which is set at time 0. Then we shall move forward one year to time $t = 1$, where we observe a spot price of €100 = 115.50, and find the contract value. We assume the contract calls for delivery of one unit of the underlying.

Case 1. No cash flows or costs on the underlying

The forward price is easily found as follows:

$$F(0, T) = S_0 (1 + r)^T$$

$$F(0, 3) = 100 \times 1.04^3$$

$$= 112.49$$

Now, moving forward to time 1, the forward contract value is found as

$$\begin{aligned}
 V_t(0, T) &= S_t - F(0, T)(1+r)^{-(T-t)} \\
 V_1(0, 3) &= S_1 - F(0, 3)(1+r)^{-(3-1)} \\
 &= 115.50 - 112.49 \times 1.04^{-2} \\
 &= 11.50
 \end{aligned}$$

Case 2. Cash flows on the underlying

Now let us assume that the underlying pays cash (dividends or interest) of €2.50 at times 1, 2 and 3. We must first find the present value at time 0:

$$C(0, T) = C(0, 3) = 2.50 \times 1.04^{-1} + 2.50 \times 1.04^{-2} + 2.50 \times 1.04^{-3} = 6.94$$

The forward price would be

$$\begin{aligned}
 F(0, T) &= (S_0 - C(0, T))(1+r)^T \\
 F(0, 3) &= (S_0 - C(0, 3))(1+r)^3 \\
 &= (100 - 6.94) \times 1.04^3 \\
 &= 104.68
 \end{aligned}$$

To obtain the value at time 3, we must find the new present value of the cash flows as of time 1.

Assuming the cash flow at time 1 is already paid, we obtain $C(1, 3)$ as

$$C(t, T) = C(1, 3) = 2.50 \times 1.04^{-1} + 2.50 \times 1.04^{-2} = 4.72$$

The value of the forward contract at time 1 is, therefore,

$$\begin{aligned}
 V_t(0, T) &= S_t - C(t, T) - F(0, T)(1+r)^{-(T-t)} \\
 V_1(0, 3) &= S_1 - C(1, 3) - F(0, 3)(1+r)^{-(3-1)} \\
 &= 115.50 - 4.72 - 104.68 \times 1.04^{-2} \\
 &= 14.00
 \end{aligned}$$

Case 3. Underlying is a currency

Let the foreign risk-free rate be $r_f = 0.03$. The forward price would be

$$\begin{aligned}
 F(0, T) &= S_0(1+r_f)^{-T}(1+r)^T \\
 F(0, 3) &= S_0(1+r_f)^{-3}(1+r)^3 \\
 &= 100 \times 1.03^{-3} \times 1.04^3 \\
 &= 102.94
 \end{aligned}$$

At time 1, the value of the contract would be

$$\begin{aligned}
 V_t(0, T) &= S_t(1 + r_f)^{-(T-t)} - F(0, T)(1 + r)^{-(T-t)} \\
 V_1(0, 3) &= S_1(1 + r_f)^{-(3-1)} - F(0, 3)(1 + r)^{-(3-1)} \\
 &= 115.50 \times 1.03^{-2} - 102.94 \times 1.04^{-2} \\
 &= 13.70
 \end{aligned}$$

Case 4. Underlying is a commodity with storage costs

Assume that the underlying is a commodity that incurs costs of €5 at the end of each year. We can assume that these costs are net of any convenience yield. The present value of the storage costs is

$$\gamma(0, 3) = 5 \times 1.04^{-1} + 5 \times 1.04^{-2} + 5 \times 1.04^{-3} = 13.88$$

The forward price would be

$$\begin{aligned}
 F(0, T) &= (S_0 + \gamma(0, T))(1 + r)^T \\
 F(0, 3) &= (S_0 + \gamma(0, 3))(1 + r)^3 \\
 &= (100 + 13.88) \times 1.04^3 \\
 &= 128.10
 \end{aligned}$$

Now move forward to time 1. The value of the storage costs is

$$\gamma(1, 3) = 5 \times 1.04^{-1} + 5 \times 1.04^{-2} = 9.43$$

The value of the forward contract is, therefore,

$$\begin{aligned}
 V_t(0, T) &= S_t + \gamma(t, T) - F(0, T)(1 + r)^{-(T-t)} \\
 V_1(0, 3) &= S_1 + \gamma(1, 3) - F(0, 3)(1 + r)^{-(3-1)} \\
 &= 115.50 + 9.43 - 128.10 \times 1.04^{-2} \\
 &= 6.49
 \end{aligned}$$

To recap, we have examined the pricing and valuation of forward contracts when the underlying is an asset. We saw that the forward price is the future value of the spot price after a deduction for any benefits received on the asset or an increment for any costs incurred on the asset. The value of a forward contract is always zero today, the spot price minus the forward price at expiration, and the spot price adjusted for costs and benefits minus the present value of the forward price any time during the life of the contract.

I.A.7.3 Principles of Pricing and Valuation for Forward Contracts on Interest Rates

Forward contracts on interest rates are widely used by corporations and financial institutions. These contracts are known as *forward rate agreements (FRAs)*. The general notion of pricing and valuation is the same as for contracts on assets, but the steps required to obtain the results are slightly more complex. First we describe some basic characteristics of FRAs.

An FRA is an agreement between two parties in which one party, the buyer or the long, agrees to make a known interest payment to the other party, the seller or the short, at a future date, with the seller agreeing to make an interest payment to the buyer based on an unknown rate that will be determined when the contract expires. Most FRAs are based on well-established interest rates such as dollar LIBOR or Euribor. The underlying instrument is a Eurodollar time deposit, a loan in which the borrower is a high-quality (but not default-free) bank. We shall simply use the term LIBOR in reference to the underlying rate. An FRA expires in a certain number of months and the underlying LIBOR is based on a certain number of months. Traders use terminology such as a '6 × 9 FRA' to describe an FRA that expires in six months with the underlying being three-month LIBOR. Hence, the '9' refers to the maturity of the underlying Eurodollar time deposit at the time the contract is created. Six months later when the FRA expires, it is a three-month time deposit.

The computations require the determination of interest payments and typically rely on the use of a day count. Let us now redefine our time framework. The FRA is created on day 0 and expires on day m . The underlying is q -day LIBOR. Let $L_0(m)$ be m -day LIBOR on day 0, $L_0(m + q)$ be $(m+q)$ -day LIBOR on day 0, and $L_m(q)$ be q -day LIBOR on day m . $L_m(q)$ is the uncertain rate revealed on day m that determines the payment from seller to buyer and should be viewed as the underlying. The fixed payment made by buyer to seller will be denoted as $F(0,m,q)$, representing the payment determined at day 0 for a contract expiring on day m in which the underlying is q -day LIBOR. We shall also be interested in the value of the FRA at a particular day during its life, which we denote as day d .

As in any interest payment, there must be an underlying principal amount on which the interest is based. This principal amount is called the *notional principal*. Notional principal is never paid because it serves no purpose other than to determine the interest payments.⁴ Our calculations for the value of the FRA apply to the case of a notional principal of one unit of the currency. If the

⁴ If notional principal were paid, the parties would merely exchange the same sum of money at the start of the contract and at the end, which would serve no economic purpose and merely create credit risk where none existed.

notional principal is N units of the currency, we must multiply the value obtained by N . In addition, the results are derived from the perspective of the party going long. To obtain the value for the party going short, we simply multiply the value for the party going long by minus one. In accordance with standard interest calculations, we adjust by a factor to reflect the period of the underlying. In the LIBOR market, an interest payment at q -day LIBOR would be obtained by multiplying the rate times days/360.

Finally, let us establish the symbols for the values of the FRA at various times during its life. Let $VFRA_0(0,m,q)$ be the value on day 0 of an FRA established at time 0, expiring at time m , with the underlying being q -day LIBOR. Let $VFRA_d(0,m,q)$ be the value on day d of an FRA established at time 0, expiring at time m , with the underlying being q -day LIBOR. Let $VFRA_m(0,m,q)$ be the value on day m (expiration) of an FRA established at time 0, expiring at time m , with the underlying being q -day LIBOR. Now we can proceed to determine the value and pricing of an FRA.

I.A.7.3.1 The Value of an FRA at Expiration

The value of an FRA at expiration is a standard calculation widely used in the industry:

$$VFRA_m(0,m,q) = \frac{(L_m(q) - F(0,m,q)) \frac{q}{360}}{1 + L_m(q) \frac{q}{360}} \quad (\text{I.A.7.12})$$

The term $L_m(q) - F(0,m,q)$ is the difference between the underlying q -day rate determined on day m , $L_m(q)$, and the rate the parties agreed to on day 0, $F(0,m,q)$. Given that all other terms in equation (I.A.7.12) are positive, the sign of the equation, which is the value of the FRA to the long, is positive if $L_m(q)$ exceeds $F(0,m,q)$. In that case, the long receives the payment from the short. If $L_m(q)$ is less than $F(0,m,q)$, equation (I.A.7.12) is negative, the value of the FRA to the long is negative, and the payment flows from the short to the long. The denominator to the equation is an adjustment factor that reflects the fact that the interest rate $L_m(q)$ is determined on day m in the LIBOR market. In that market, $L_m(q)$ is the rate on a Eurodollar time deposit that begins on day m and pays the interest q days later. But in the FRA market, the payment at the rate $L_m(q)$ is determined on day m and paid on day m . Thus, in a sense, this payment is somewhat premature and, accordingly, is discounted at q -day LIBOR on day m . Naturally all rates are adjusted by the factor $q/360$.

So, at the expiration date, the parties exchange the net difference between interest rates adjusted by the factor $q/360$ and discounted to reflect the fact that the payment is received earlier than implied by the underlying rate.

I.A.7.3.2 The Value of an FRA at the Start

At the start of the contract, the value of the FRA has to be zero, because no money changes hands.

$$V_{FRA_0}(0,m,q) = 0 \quad (\text{I.A.7.13})$$

Of course, this is the same result we obtained for forward contracts on assets.

I.A.7.3.3 The Value of an FRA During Its Life

To value an FRA during its life, we use the same general approach as when we were valuing a forward contract on an asset: we position ourselves during the life of the contract and engage in a transaction that will produce the same value of the contract at expiration. So we now put ourselves on day d . The information available to us is $(m - d)$ -day LIBOR, denoted as $L_d(m - d)$, and $(m + q - d)$ -day LIBOR, denoted as $L_d(m + q - d)$.

To replicate the value of the FRA at its expiration, at time d we invest the following amount in a LIBOR time deposit:

$$\frac{1}{1 + L_d(m - d)\left(\frac{m - d}{360}\right)}$$

This instrument will pay \$1 on day m . We also issue a LIBOR time deposit promising to pay $1 + F(0,m,q)(q/360)$ on day $m + q$. This loan will produce upfront cash of

$$\frac{1 + F(0, m, q)\left(\frac{q}{360}\right)}{1 + L_d(m + q - d)\left(\frac{m + q - d}{360}\right)}$$

On day m , the time deposit we hold is worth \$1. The loan has a value of

$$-\left(\frac{1 + F(0, m, q)\left(\frac{q}{360}\right)}{1 + L_m(q)\left(\frac{q}{360}\right)}\right)$$

The total value of the strategy on day m is, therefore,

$$1 - \left(\frac{1 + F(0, m, q) \left(\frac{q}{360} \right)}{1 + L_m(q) \left(\frac{q}{360} \right)} \right)$$

Using a common denominator, the value of the strategy on day d is

$$\begin{aligned} & \frac{1 + L_m(q) \left(\frac{q}{360} \right) - \left(1 + F(0, m, q) \left(\frac{q}{360} \right) \right)}{1 + L_m(q) \left(\frac{q}{360} \right)} \\ &= \frac{(L_m(q) - F(0, m, q)) \left(\frac{q}{360} \right)}{1 + L_m(q) \left(\frac{q}{360} \right)} \end{aligned}$$

which is the value of the FRA, equation (I.A.7.12), at expiration. Thus, on day d , this strategy replicates the FRA. So its value on day d must be the same as the value of the FRA on day d :

$$VFRA_d(0, m, q) = \frac{1}{1 + L_d(m - d) \left(\frac{m - d}{360} \right)} - \frac{1 + F(0, m, q) \left(\frac{q}{360} \right)}{1 + L_d(m + q - d) \left(\frac{m + q - d}{360} \right)} \quad (\text{I.A.7.14})$$

I.A.7.3.4 Pricing the FRA on Day 0

We know that the value of the FRA on day 0 is zero. We also know that the value of the FRA at expiration can be replicated by pursuing the strategy described in the previous section on day d . Now let day d be day 0, restate equation (I.A.7.14) with $d = 0$, set the equation to zero, and solve for $F(0, m, q)$ to obtain:

$$F(0, m, q) = \left(\frac{1 + L_0(m + q) \left(\frac{m + q}{360} \right)}{1 + L_0(m) \left(\frac{m}{360} \right)} - 1 \right) \left(\frac{360}{q} \right) \quad (\text{I.A.7.15})$$

Although this expression looks complex, it is actually quite simple. The numerator of the first fraction, $1 + L_0(m + q)((m + q)/360)$, contains the compound value of a \$1 LIBOR time deposit of $m + q$ days. The denominator, $1 + L_0(m)(m/360)$, contains the compound value of a \$1 LIBOR time deposit of m days. The division of the former by the latter, after subtracting 1, yields

the compound future value of a \$1 LIBOR forward time deposit, initiated on day m and maturing on day $m + q$. The second fraction, $360/q$, simply annualises this rate. Thus, the fixed rate on the FRA is forward LIBOR.

I.A.7.3.5 Numerical Examples

Now let us illustrate these formulae for an FRA. Consider a 3×9 FRA, meaning that the contract expires in three months (90 days) and the underlying is six-month (180-day) LIBOR. Assume at time 0, the spot rate is 8% for 90-day LIBOR and 8.5% for 270-day LIBOR. Thus, our information is as follows: $m = 90$, $q = 180$; $L_0(m) = L_0(90) = 0.08$; $L_0(m + q) = L_0(270) = 0.085$.

The forward price is

$$\begin{aligned}
 F(0, m, q) &= \left(\frac{1 + L_0(m + q) \left(\frac{m + q}{360} \right)}{1 + L_0(m) \left(\frac{m}{360} \right)} - 1 \right) \left(\frac{360}{q} \right) \\
 F(0, 90, 180) &= \left(\frac{1 + L_0(270) \left(\frac{270}{360} \right)}{1 + L_0(90) \left(\frac{90}{360} \right)} - 1 \right) \left(\frac{360}{180} \right) \\
 &= \left(\frac{1 + .085 \left(\frac{270}{360} \right)}{1 + .08 \left(\frac{90}{360} \right)} - 1 \right) \left(\frac{360}{180} \right) \\
 &= 0.0858
 \end{aligned}$$

Now move 60 days forward. Let the 30-day rate be 9.2% and the 210-day rate be 9%. The information we have is $d = 60$, $m - d = 90 - 60 = 30$, $m + q - d = 90 + 180 - 60 = 210$; $L_d(m - d) = L_{60}(30) = 0.092$; $L_d(m + q - d) = L_{60}(210) = 0.09$. The value of the FRA per \$1 would be

$$\begin{aligned}
 VFRA_d(0, m, q) &= \frac{1}{1 + L_d(m - d) \left(\frac{m - d}{360} \right)} - \frac{1 + F(0, m, q) \left(\frac{q}{360} \right)}{1 + L_d(m + q - d) \left(\frac{m + q - d}{360} \right)} \\
 VFRA_{60}(0, 90, 180) &= \frac{1}{1 + .092 \left(\frac{30}{360} \right)} - \frac{1 + .0858 \left(\frac{180}{360} \right)}{1 + .09 \left(\frac{210}{360} \right)} \\
 &= 0.0015
 \end{aligned}$$

I.A.7.4 The Relationship Between Forward and Futures Prices

Futures contracts are similar to forward contracts but are standardised, guaranteed against default, and trade in a regulated environment on a futures exchange. Any observed difference between forward and futures prices is typically attributed to the different cash flow streams of the two contracts. Futures provide cash flows daily in the form of the daily settlement of gains and losses, while forward contracts provide cash flows at the end of the life of the contract.

The differences between forward and futures prices for assets have been examined by Cox et al. (1981) and Jarrow and Oldfield (1981). Summarising their findings, we can say that forward and futures prices for assets are the same at expiration, or if the contracts have a single day to go before expiration, or if certain relationships regarding the correlation between futures prices, forward prices and risk-free bond prices are met. The last condition is best understood by considering one party holding a forward contract and another holding a futures contract. If futures prices and interest rates are positively related, the party holding the futures contract will have an advantage over the party holding the forward contract because gains from daily settlement will be invested at rising interest rates and losses from daily settlement will be funded at falling interest rates. The stream of interest rates over the life of the contract will have no effect on the value of the forward contract at expiration. If futures prices and interest rates are negatively related, the party holding the futures contract will have a disadvantage over the party holding the forward contract because gains from daily settlement will be invested at falling interest rates and losses from daily settlement will be funded at rising interest rates. The party holding the advantage will be willing to pay a higher price. If futures price and interest rates are unrelated, neither contract will have an advantage, and forward and futures prices will be equal.⁵

The difference between forward and futures prices for FRAs relative to Eurodollar futures is much less clear. A difference does exist and has been examined by Sundaresan (1991), but the issues are beyond this level of treatment. Credit risk can make futures prices differ from forward prices, but any such difference will be attributed to a complex mixture of factors such as which party has the better credit risk and the existence of credit mitigation factors for forward contracts.⁶

⁵ There are a couple of other conditions as well as other ways of stating the necessary conditions for a divergence between futures and forward prices, but we do not pursue these here.

⁶ In fact, parties to some forward contracts engage in periodic settlements, rendering the cash flow streams of these contracts more like those of futures contracts.

References

Cox, JC, Ingersoll, JE, Jr, and Ross, SA (1981) 'The relation between forward prices and futures prices', *Journal of Financial Economics*, 9 (4), pp. 321–46.

Jarrow, R, and Oldfield, G (1981) 'Forward contracts and futures contracts', *Journal of Financial Economics*, 9 (4), pp. 373–82.

Sundaresan, S (1991) 'Futures prices on yields, forward prices, and implied forward prices from term structure', *Journal of Financial and Quantitative Analysis*, 26 (3), pp. 409–24.

I.A.8 Basic Principles of Option Pricing

Paul Wilmott

I.A.8.1 Factors Affecting Option Prices

Option values clearly depend on the value of the underlying stock. If you own a call option giving you the right to buy the stock at a fixed price, then you benefit if the stock price rises. Fingers crossed and touching wood, if the stock goes up you make a tidy profit, but if the stock falls you could lose all of your investment. We can safely say that the option value depends on the stock price and, when we come to the mathematics, we can say that the option value is a *function* of the stock price.

Option values depend on many more quantities than just the stock price. The time to expiry is another important variable in the pricing problem. The dependence on time works in two ways. First, the longer the time to expiry, the more the stock price can fluctuate. This may increase or decrease the option value, depending on the structure of the option in question. Second, the longer before you get a payoff, the less it is worth: this is simple time value of money (see Chapter I.B.1).

Asset price and time are commonly referred to as *variables* in the option-pricing problem. There are also the *parameters*. There are parameters associated with the option, parameters associated with the stock and one parameter associated with the currency in which the stock is denominated. Let us take these in turn.

The parameters associated with the option will depend on the complexity of the option. However, for vanilla options the parameters are just the expiration date and the strike, and also other details of the payoff such as whether we have a call or a put, and whether the contract is European or American.

Parameters associated with the stock will depend on what model you are using for the evolution of the stock price over time. That said, in the equity markets 99% of the time we use what is called a *lognormal random walk*. This has three basic parameters: (1) the growth rate of the stock, (2) its volatility and (3) its dividends. The growth rate determines how quickly the stock price rises and the volatility measures how random the stock price is.

Intuition would tell us that the faster a stock price grows the more a call option would be worth, and the less a put. Unfortunately, this intuition is wrong (at least according to the popular pricing theories). We will see why this is so later on. (See also Wilmott, 2000, Chapters 5 and 12.)

Reassuringly, option values do depend on the stock volatility. This volatility (or annualised standard deviation of returns in statistical terms) is a very important factor in the pricing of derivatives. Option prices depend very strongly on volatility. But volatility is impossible to see. It can be estimated and forecast, but using only statistical models of volatility. In a sense this is very fortunate for option markets. As long as people can disagree on the value of volatility, there will be a market – one person selling and another buying because they both think they are getting a good deal.

The final parameters determining an option value are the interest rate of the currency in which the stock is denominated and the dividend, if the underlying is a stock, or the foreign interest rate, if the underlying is an exchange rate.

The parameters mentioned above – volatility, interest rates and dividend – are assumed to be constant in the basic Black–Scholes model. More advanced models treat them more realistically, in all manner of clever ways (Wilmott, 2000). There are other factors affecting prices, but are not usually incorporated into the simpler models, such as credit risk and transaction costs.

I.A.8.2 Put–Call Parity

Imagine that you buy one European call option with a strike of K and an expiry of T and that you write a European put option with the same strike and expiry. Today's date is t . The payoff you receive at T for the call will look like Figure I.A.8.1. The payoff for the put is the line in Figure I.A.8.2. Note that the sign of the payoff is negative, since you *wrote* the option and are liable for the payoff. The payoff for the portfolio of the two options is the sum of the individual payoffs, shown in Figure I.A.8.3. In mathematical terms, the payoff for this portfolio of options is

$$\max(S(T) - K, 0) - \max(K - S(T), 0) = S - K$$

where $S(T)$ is the value of the underlying asset at time T .

Figure I.A.8.1

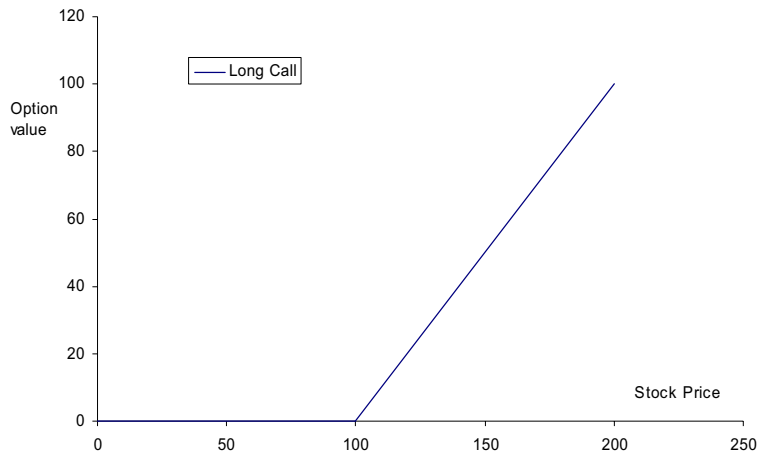


Figure I.A.8.2

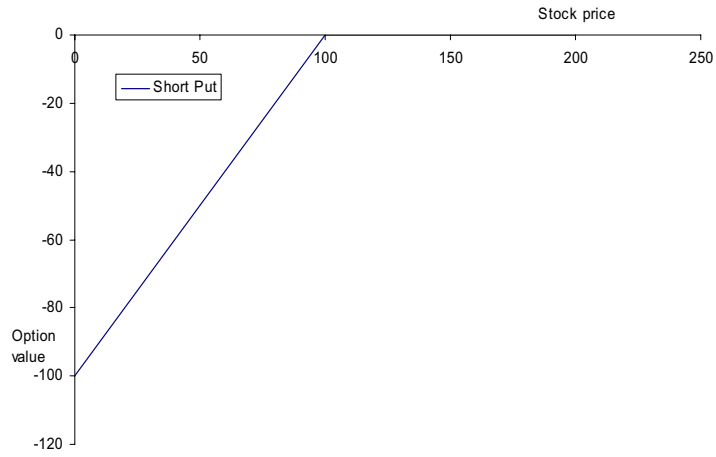
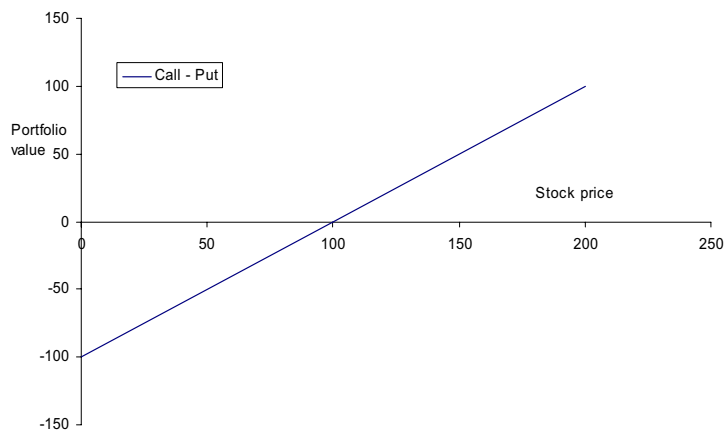


Figure I.A.8.3



The right-hand side of this expression consists of two parts, the asset and a fixed sum K . Is there another way to get exactly this payoff? Yes, there is. And it is much simpler than buying and selling options.

If I buy the asset today it will cost me $S(t)$ and be worth $S(T)$ at expiry. I don't know what the value $S(T)$ will be but I do know how to guarantee to get that amount, and that is to buy the asset now. What about the K term? To lock in a payment of K at time T involves a simple trade in a zero-coupon bond at time t . This bond must have a maturity that is the same as the expiration of the option. The conclusion is that the portfolio of a long call and a short put gives me exactly the same payoff as a long-asset, short-bond position. The equality of these cash flows is independent of the future behaviour of the stock.

Using C and P to denote values of the call and the put respectively, and Z for the value of the bond having value 1 at maturity, we can write

$$C - P = S - KZ .$$

This relationship holds at any time up to expiry and is known as *put-call parity*. If this relationship did not hold then there would be riskless arbitrage opportunities.

This relationship is very special in finance, in that it is completely independent of any mathematical model for the behaviour of the stock price. Such relationships are few and far between.

The equality also shows that one's intuition about the dependence of an option's value on the stock direction is incorrect. For if the call option were to go up in value because one expected a high payoff, then the put would also have to go up in value even though it would simultaneously have a lower chance of getting any payoff.

Dividends can easily be incorporated into the put-call parity relationship. The role of the stock price in the formula becomes the stock price less the present value of all of the dividends from now until the option expiration. Of course, this assumes that the amount of the dividend payment is already known.

I.A.8.3 One-step Binomial Model and the Riskless Portfolio

What is the simplest model that captures the essential randomness that we see in stock prices? It has to be the binomial model in which a stock can go up only a given amount, or down only a given amount, but the direction is left to chance (see Cox and Rubinstein, 1985). This very basic

model may not be perfect; however, it does contain the essential ingredients that lead to some of the most fundamental concepts in derivatives theory.

Figure I.A.8.4 shows the basics of the binomial model. It contains several parameters. These parameters are the current level of the stock, the stock price should it move up, the stock price should it move down, the probability of a rise in the stock price, and consequently the probability of a fall, and the time step between now and when the stock prices changes.

Figure I.A.8.4

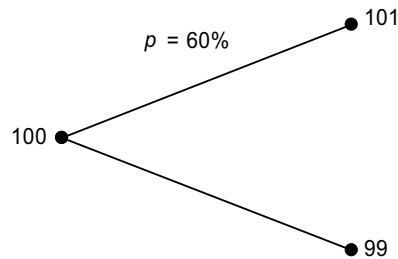


Figure I.A.8.4 contains numbers for all of these parameters: stock price of 100, going up to 101 or down to 99, etc. In practice these numbers will be linked to properties of the asset in question, such as its volatility. There is also one final parameter, the risk-free interest rate.

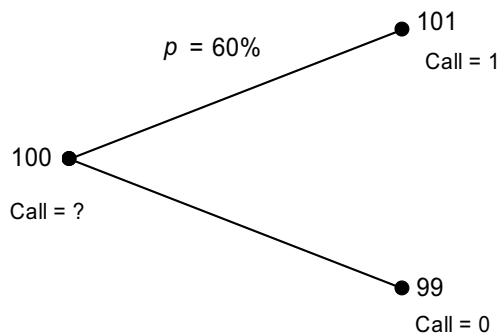
To keep the example as concrete as possible, we'll assume that the stock price move is over one day, from today to tomorrow. Now let's introduce an option on this stock, an option that expires 'tomorrow'. It'll be a call option with strike of 100, although the following ideas can be applied to options generally. We can make two obvious (and correct) statements about the value of this call option:

- if the stock price rises tomorrow to 101, the call will have a payoff of 1;
- if the stock price falls to 99, the option will expire worthless (Figure I.A.8.5).

Since the probability that the stock price will rise is 60%, can we make any more obvious statements? Could we look at the expected payoff for the option and treat that as an option 'price'? Could we say that the value of the option today, assuming interest rates are zero, is 0.6 since that is the expected payoff? No, we cannot. That statement, although intuitively appealing, is actually incorrect. We cannot price something using a simple expectation, since that would completely ignore the risk in the option payoff.

To see why that is incorrect and what is the correct price for the option, we need to introduce the idea of delta hedging.

Figure I.A.8.5



I.A.8.4 Delta Neutrality and Simple Delta Hedging

Risk reduction can be done in one of two basic ways: by exploiting the lack of correlation between instruments, which is diversification; or by exploiting a high degree of correlation between instruments, which is called hedging.

Delta hedging aims to reduce risk from stock price movements and is ubiquitous in the derivatives world. It is used universally and is one of the foundations of the theory.

Consider the option and stock that we had in the above binomial example. Now imagine a portfolio containing one call option and short one-half of the underlying stock. We do not know the value of this portfolio today since we do not yet have a value for the option. That is our goal.

Although we do not know the value of the portfolio today we do know its value tomorrow. If the stock price rises we have a portfolio made up of an option worth 1 and a short stock position worth -0.5×101 . The minus sign is because it is a short stock position and the 0.5 because we hold (short) half a share:

$$1 - \frac{1}{2} \times 101 = -\frac{99}{2}.$$

Now, what happens if the stock price falls to 99? The option expires, worthless, and so our portfolio is worth

$$0 - \frac{1}{2} \times 99 = -\frac{99}{2}.$$

The portfolio takes the same value whether the stock rises or falls. There is therefore no risk in the position. That is perfect delta hedging.

We are one step away from pricing the option.

If interest rates are zero, the $-99/2$ that we will certainly have tomorrow must be worth $-99/2$ today. If we use V to denote the unknown option value we can write

$$V - \frac{1}{2} \times 100 = -\frac{99}{2} .$$

From which we deduce that $V = 1/2$.

This binomial pricing method can be used in practice to price many types of option. There are three final pieces of information you will need before using the method on a real problem:

1. How did I know to sell 0.5 of the stock?
2. What if interest rates are not zero?
3. What if we are not one day from expiration?

Use the Greek letter Δ (delta) to denote the quantity of stock we sell short for hedging. The portfolio value if the stock rises would then be

$$1 - \Delta 101 .$$

If the stock falls, we have

$$0 - \Delta 99 .$$

To 'hedge' the option payoff with the stock we just choose Δ such that

$$1 - \Delta 101 = 0 - \Delta 99 .$$

The solution of which is

$$\Delta = \frac{1 - 0}{101 - 99} = \frac{1}{2} .$$

So we must sell 0.5 short. The above expression can be used for more general option payoffs and arbitrary stock prices. The rule is then

$$\Delta = \frac{\text{Change in Option Value}}{\text{Change in Stock Value}} .$$

In the limit as the stock range becomes smaller and smaller this expression turns into a gradient/slope/sensitivity, and we talk about the delta being the rate of change of option value with respect to stock price.

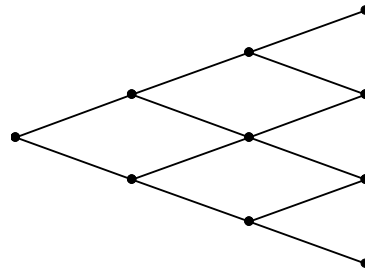
If interest rates are non-zero, this does not affect the above expression for the delta. However, when we discount back from ‘tomorrow’ to ‘today’ we must multiply by a suitable discount factor. In the example, $-99/2$ tomorrow would be worth

$$-\frac{99}{2} \times \text{DF},$$

where DF is a discount factor, less than 1. Otherwise the option-pricing principle is the same.

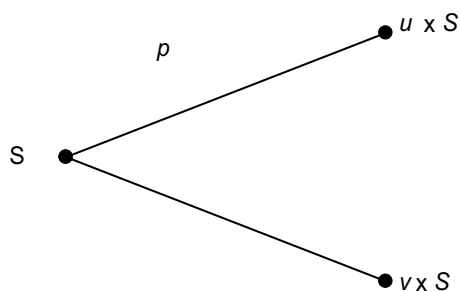
Finally, if we are so far from expiration that a simple up/down model will not be good enough, we need to break time until expiration up into several time steps. This is illustrated in Figure I.A.8.6. The stock price ‘tree’ is built up from today’s stock price first of all. Then we treat each one-step tree as in the above example and work from right to left across the tree, filling in the option prices as we go.

Figure I.A.8.6



In the above example we had a stock price that was to take one of two values a day later. Using more general notation, we have a situation like that shown in Figure I.A.8.7.

Figure I.A.8.7



The stock price is currently S , can rise to uS or fall to vS . The probability of the rise is p . In our example $S = 100$, $u = 1.01$, $v = 0.99$ and $p = 0.6$. Now let us do the above analysis, using symbols, in order to generalise the binomial model. There are several key steps.

1. Pick u and v . The choice of these is governed by the volatility of the underlying asset and the time step.
2. Set up the hedged portfolio. At this stage you will not know how much of the underlying asset to hedge with, so call the quantity held short Δ .
3. Choose Δ such that the portfolio values at expiry are the same, whether the asset moves up or down.
4. Discount this portfolio value to the present and calculate the current option value.

The three constants u and v are chosen to give the binomial model the same characteristics as the asset we are modelling. That means we want to capture the volatility of the underlying asset, in particular the volatility over the next time step, dt . Although we will not be needing it, we will also carry around the probability p for a while. We choose

$$u = 1 + \sigma\sqrt{dt} ,$$

$$v = 1 - \sigma\sqrt{dt}$$

and

$$p = \frac{1}{2} + \frac{\mu\sqrt{dt}}{2\sigma} .$$

We have introduced two new parameters here: μ , the drift of the asset; and σ , the volatility. We have chosen these parameters in such a way that (approximately)

- the expected return on the asset is μdt ,
- the standard deviation of returns is $\sigma dt^{1/2}$.

Suppose we are one time step away from expiry. Construct a portfolio consisting of one option and a short position in a quantity Δ of the underlying. This portfolio has value $\Pi = V - \Delta S$, where the value of the option V is to be determined. One time step later, at expiry, the portfolio takes one of two values, depending on whether the asset rises or falls. These two values are

$$V^+ = -\Delta uS \quad \text{and} \quad V^- = -\Delta vS.$$

V^+ is the option value if the asset rises and V^- the value if it falls. The values of both of these expressions are known if we know Δ . And Δ is under our control. Having the freedom to choose Δ , we can make the value of this portfolio the same whether the asset rises or falls. This is ensured if we make

$$\Delta = \frac{V^+ - V^-}{(u - v)S}.$$

Then the new portfolio value is

$$\frac{uV^- - vV^+}{u - v}.$$

Since the value of the portfolio has been guaranteed, we can say that its value must coincide with the value of the original portfolio plus any interest earned at the risk-free rate, r . Thus

$$(1 + r dt) \left(V - \frac{V^+ - V^-}{u - v} \right) = \frac{uV^- - vV^+}{u - v}.$$

Rearranging, we get

$$(1 + rdt)V = p'V^+ + (1 - p')V^- \tag{I.A.8.1}$$

where

$$p' = \frac{1}{2} + \frac{r\sqrt{dt}}{2\sigma}.$$

The right-hand side of equation (I.A.8.1) is just like an expectation: it is the sum of probabilities multiplied by outcomes. Let us compare the expression for p' with the expression for the actual probability p . The two expressions differ in that where one has the interest rate r the other has the drift μ , but they are otherwise the same.

We see that the probability of a rise or fall is irrelevant as far as option pricing is concerned. But what if we interpret p' as a probability? Then we could say that the option price is the present value of an expectation. But not the ‘real’ expectation. We will come on to the meaning of ‘real’

in this context shortly. We call p' the *risk-neutral probability*. It is like the binomial probability when the drift rate is r instead of μ .

Observe that the risk-free interest plays two roles in option valuation. It is used once for discounting to give present value, and it is used as the drift rate in the asset price random walk.

Example I.A.8.1: Pricing an option with the binomial model

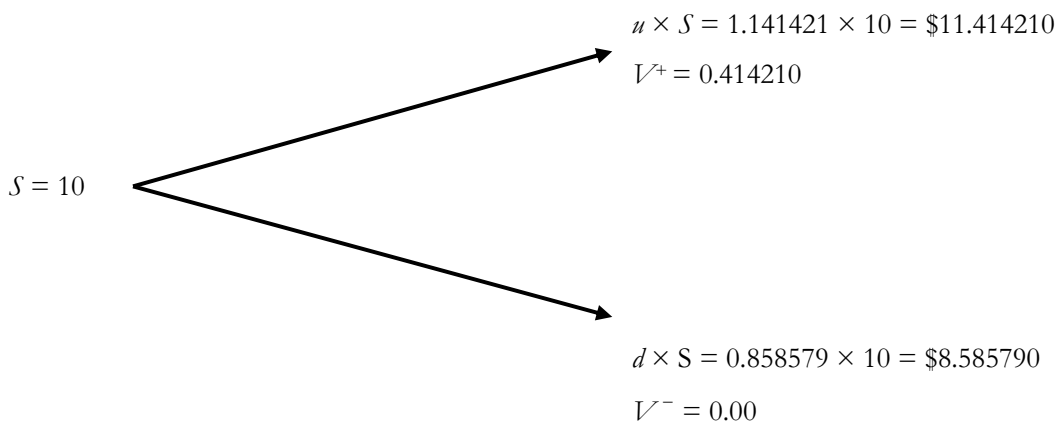
A European call option over a certain stock has a term of six months (or 0.5 years). The current stock price is \$10.00 and the strike of the option is \$11.00. The risk-free rate is 4.0% p.a. The volatility of the stock is 20% p.a. *What is the value of the option using a one-step binomial model?* We start by building a stock price tree using the equations for u and v that were given earlier:

$$u = 1 + \sigma\sqrt{dt} \quad \text{and} \quad v = 1 - \sigma\sqrt{dt} .$$

Substituting values for σ (0.2) and dt (0.5) we have:

$$u = 1 + 0.2\sqrt{0.5} = 1.141421 \quad \text{and} \quad v = 1 - 0.2\sqrt{0.5} = 0.858579 .$$

Using these values for u and v the stock price tree is as follows:



At the upper terminal node the value of the stock is \$11.414210. Here the call option would be exercised to give a payoff of $S(T) - K$ or \$0.414210, therefore $V^+ = \$0.414210$. At the lower terminal node the value of the stock is \$8.585790. Here the call option would not be exercised so $V^- = \$0.00$.

We use equation (I.A.8.1) to determine V the value of the option today, but first we must calculate p' , the risk-neutral probability that the stock will increase:

$$p' = \frac{1}{2} + \frac{r\sqrt{dt}}{2\sigma} = \frac{1}{2} + \frac{0.04\sqrt{0.5}}{2 \times 0.2} = 0.570711$$

Substituting all these values into equation (I.A.8.1), we can solve for V :

$$(1 + (0.04 \times 0.5))V = (0.570711 \times 0.414210) + ((1 - 0.570711) \times 0) = 0.236394$$

giving $V = \$0.2318$.

How would the value of the option differ for a European put option, assuming all other facts the same?

The valuation in the case of a put option proceeds in exactly the same way. A difference in the option values arises because of the way the payoff is defined for a put versus a call. For the upper terminal node the stock price is at \$11.414210. A put option with a strike price of \$10.00 would not be exercised, so $V^+ = 0.0$. For the lower terminal node the stock price is at \$8.585790. The payoff of the put option is $K - S(T) = 11.00 - 8.585790$, therefore $V^- = \$2.414210$. Substituting values into equation (I.A.8.1), we can solve for V :

$$(1 + (0.04 \times 0.5))V = (0.570711 \times 0.0) + ((1 - 0.570711) \times 2.414210) = 1.036394$$

giving $V = \$1.0161$.

I.A.8.5 Risk-Neutral Valuation

The binomial tree algorithm is a completely deterministic method for pricing options. But it also has an interpretation that is entirely probabilistic. I said above that expectations were irrelevant as far as pricing is concerned, but that is not entirely true. There is a simple interpretation of the binomial algorithm that uses expectations. Interpreting p' as a probability, the option-pricing equation is the statement that *the option value at any time is the present value of the risk-neutral expected value at any later time*.

You can think of an option value as being the present value of an expectation, but it is not the real-world expectation. This can be confusing; when it comes to pricing derivatives the real

probability of a stock price rising or falling does not matter. However, there is a fictitious probability, called the *risk-neutral probability*, that is relevant.

This is a point that cannot be stressed too much. Be careful how you calculate expectations when pricing options, remember this is where we came in. Our original question about the price of an option in the binomial setting was deliberately designed to trap you, to emphasise the point that real-world expectations are not used in pricing.

The above model and analysis has introduced several key concepts:

- There is only one price for an option, and that does not depend on which direction you think the asset price is going, only on its volatility.
- That option price can be interpreted as an expectation. This is why we can actually value options by using simulation/Monte Carlo methodology.
- There is a simple algorithm for calculating an option price a time step before expiry. This can be extended to value an option at any time before expiry.

I.A.8.6 Real versus Risk-Neutral

Everyone gets confused by the concept of risk-neutral at first, so do not feel bad if this is your first time. ‘Real’ simply means anything pertaining to the actual random walk and probabilities. In our binomial example the real probability of the stock price rising is 60%. If you were to follow this stock in an infinite number of parallel universes you would see it rising 60% of the time. ‘Risk-neutral’ (less simply) means anything pertaining to a the world in which all assets have an expected return that is the same return as money in the bank. In our binomial example the risk-neutral probability of the stock rising is 50%. That is because, if the stock were to rise with a 50% probability, the average return would be zero, since we assumed in the example that interest rates were 0%. If you were to follow this stock in an infinite number of parallel universes you would see it rising – 60% of the time, not 50%.

Once we know what the risk-neutral probabilities are, they can be used for pricing options.

Note the following facts about real and risk-neutral:

- The risk-neutral world does not exist. *True.*
- Real probabilities do not affect option prices (at least not in the common theories). *True.*
- Risk-neutral probabilities are used in the pricing of options. *True.*

Note also the following fallacies:

- People are risk-neutral, they don't care about risk. *False*. People are generally risk-averse, in some vague, inconsistent and unquantified way.
- Option prices are given by the present value of the expected payoff. *False* – unless you squeeze in the adjective 'risk-neutral' in front of expectation.
- Options are riskless because we price them in a risk-neutral world. *False* – unless you are hedging them, as explained above.

I.A.8.7 The Black–Scholes–Merton Pricing Formula

The binomial model is excellent for explaining the concept of delta hedging. However, it does suffer from being quite restricted in its representation of stock price movements. The use of more sophisticated tools, such as stochastic calculus, has proved to be an even better foundation for the theory and practice of option pricing and hedging. These tools can be used to derive the famous Black–Scholes (and Merton) or BSM option-pricing model and formulae. On a historical note, it should be pointed out that the famous pricing formulae were actually first derived, although never publicly published, by Ed Thorp and Sheen Kassouf (see Kassouf and Thorp, 1967; Tudball, 2001). For more option-pricing formulae, see Haug (1998).

The BSM option-pricing model (see Black and Scholes, 1973; Merton, 1973) is based on the same principles of delta hedging and no arbitrage but uses different mathematics. In particular, it uses a continuous-time, continuous-asset model, known as the *lognormal model*, and the techniques of stochastic calculus.

We will not go into the details here, but the model results in a partial differential equation for the price of an option. (Actually, you get the same equation if you start with the binomial algorithm and let the number of time steps increase to infinity.) This differential equation can be solved in some special cases, and under certain assumptions. We will look at the major assumptions first, and then at the resulting solutions.

The main assumptions are as follows:

- The underlying asset follows a lognormal random walk (based on Brownian motion).
- The volatility is constant and known.
- The interest rate is a known constant.
- There are no dividends on the underlying.
- The stock can be sold short.

- Any quantity of the stock may be bought or sold, it does not have to be a round number.
- There are no costs associated with buying and selling the stock.
- There is no risk of default.
- Early exercise of the option is not permitted (the option is European).

With these assumptions, the value of a call option is

$$V = SN(d_1) - Ke^{-r(T-t)}N(d_2)$$

where S is the asset price, E the strike, t the current time, and T the expiration, and r the risk-free interest rate. Also

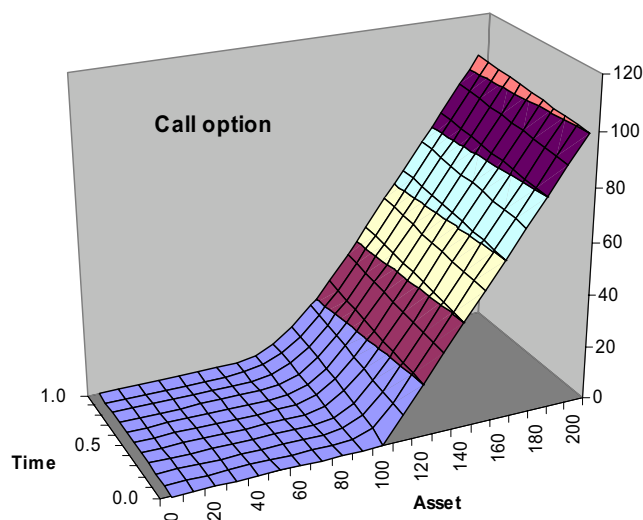
$$d_1 = \frac{\ln(S/K) + \left(r + \frac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{T-t}} \quad \text{and} \quad d_2 = d_1 - \sigma\sqrt{T-t}$$

and $N(\cdot)$ is the cumulative distribution function for the normal distribution, namely

$$N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp(-s^2/2) ds.$$

The appearance of the cumulative distribution function for the normal distribution adds weight to the probabilistic interpretation of an option's value as mentioned above.

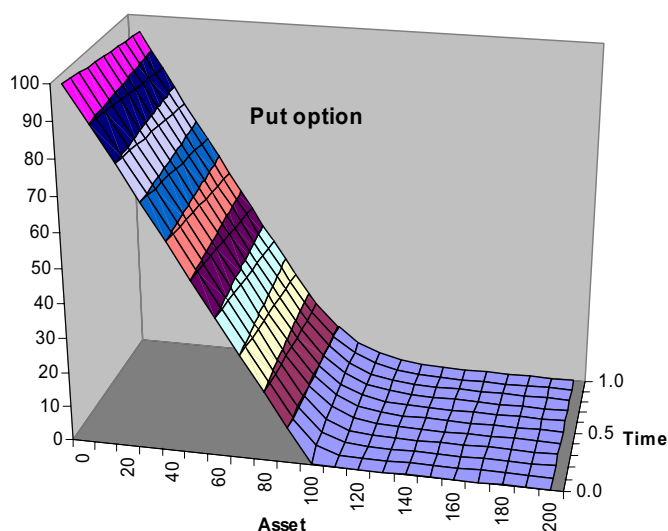
Figure I.A.8.8



The put option formula is similarly

$$V = -SN(-d_1) + Ke^{-r(T-t)}N(-d_2).$$

Figure I.A.8.9



Example I.A.8.2: Option pricing using the BSM model

For purposes of comparison we use the set of facts that were considered in Example 1. A stock is priced at \$10.00 and has volatility of 20% p.a. A European call option over the stock has a term of six months (0.5 years) and a strike of \$11.00. Value the option assuming that the risk-free rate is 3.92% p.a., quoted as a continuous rate (see I.B.1).¹

Substituting the values above into the equation for d_1 we have:

$$d_1 = \frac{\ln(10.00 / 11.00) + \left(0.0392 + \frac{1}{2} 0.20^2\right) 0.5}{0.2\sqrt{0.5}} = -0.464641$$

and

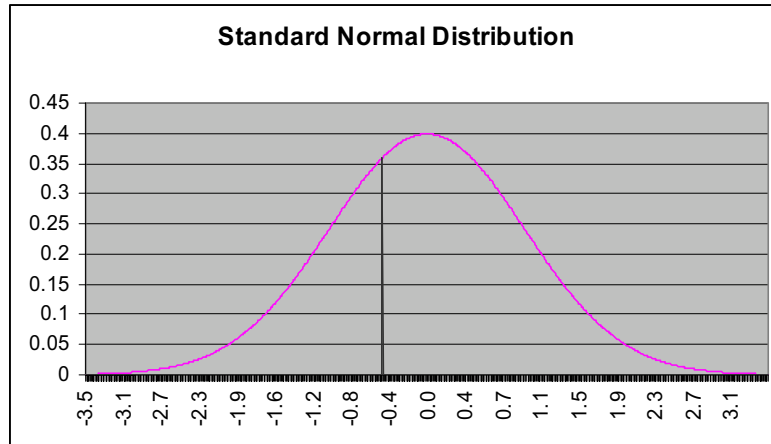
$$d_2 = 0.464641 - 0.2\sqrt{0.5} = -0.606062$$

The calculation of $N(d_1)$ and $N(d_2)$ can be done easily using standard normal tables or using the preprogrammed functions in Excel or other spreadsheet packages. Figure I.A.8.10 shows the standard normal distribution, with a vertical line at -0.464641 (the value of d_1 that concerns us in this example). $N(d_1)$ is the area under the normal curve to the left of -0.464641 , which is

¹ Note that the continuous risk-free rate used here is equivalent to the simple interest rate (of 4% p.a.) used in Example 1.

0.391024 in this case. That is, approximately 39% of the area under the standard normal curve lies to the left of -0.464641 . Further discussion of the normal distribution may be found in Chapter II.B.

Figure I.A.8.10



Similarly, $N(d_2)$ or $N(-0.606062)$ is equal to 0.272237. We can then value the call option:

$$V = SN(d_1) - Ke^{-r(T-t)}N(d_2)$$

So $V = (10 \times 0.391024) - (11e^{-0.0392 \times 0.5} \times 0.272237) = 0.2744$. The value of a European put option (other facts constant) would be:

$$V = -SN(-d_1) + Ke^{-r(T-t)}N(-d_2)$$

where $N(-d_1) = 0.678906$ and $N(-d_2) = 0.727763$. So $V = (-10 \times 0.678906) + (11e^{-0.0392 \times 0.5} \times 0.727763) = 1.0610$.

Comparing the results of the pricing calculations in Examples I.A.8.1 and I.A.8.2, we observe that the option values are not identical but are broadly similar (see Table I.A.8.1).

The similarity between the two pricing models comes about because the two make very similar assumptions about the underlying price process. The difference arises because the BSM model is in continuous time but the binomial model here uses one period of discrete time. If a multistep binomial model were used, the option values would (in most cases) be even closer.

Table I.A.8.1: Comparison of pricing approaches

	One-step binomial model	BSM model
European call	0.2318	0.2744
European put	1.0161	1.0610

I.A.8.8 The Greeks

There is a great deal more to options than pricing. In fact, it is probably more important to know how to risk manage an option than to price it. To that end people measure various quantities called *sensitivities* of an option price. These quantities are called the Greeks because they are represented by Greek letters: for instance delta, gamma, theta, vega² and rho.

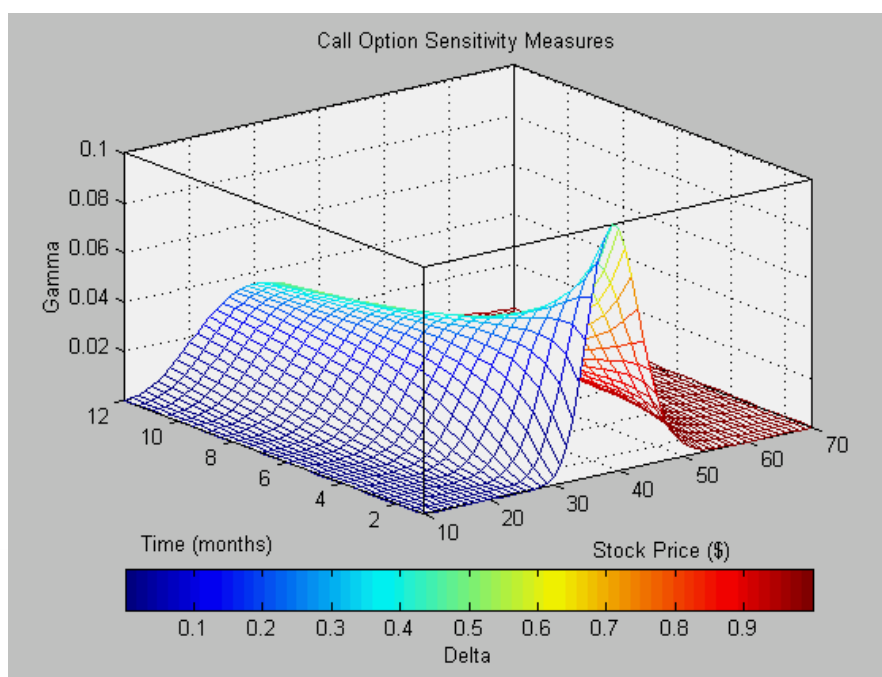
The Greeks are defined as *partial derivatives* of the option price function. More details can be given once the concept of a derivative has been introduced. Section II.C.4.3 provides a more formal introduction to the Greeks. The Greeks are used to hedge the risk of an option, or a portfolio of options. We show exactly how this is done in Section II.D.2.3

The *delta*, Δ , of an option or a portfolio of options is the sensitivity of the option or portfolio to the underlying. It is the rate of change of value with respect to the underlying.³ We have already seen the delta in Section I.A.8.4, and how it is used for hedging purposes.

The *gamma*, Γ , of an option or a portfolio of options is the second derivative of the position with respect to the underlying, or equivalently the derivative of the delta with respect to the stock price. Since gamma is the sensitivity of the delta to the underlying it is a measure of by how much or how often a position must be re-hedged in order to maintain a delta-neutral position. An option that is both close to expiry and at-the-money typically has a large (absolute) gamma; gamma may be positive or negative, depending on whether you are short or long the option. An option with a large (absolute) gamma is more sensitive to market movements; they will necessitate large adjustments to the hedging portfolio as delta changes. An option or portfolio of options with long (short) gamma will profit (incur losses) from market gaps, even if the trader is endeavouring to maintain a delta-neutral position. The only way to hedge this gamma risk is by trading another option position with an equal and opposite gamma.

² Confusingly, vega is actually not a Greek letter.

³ Note that *the underlying should always be the hedging instrument*. For instance, with an equity index option the underlying may be the index future, assuming you cannot buy the index itself.

Figure I.A.8.11: The Gamma of a European Call Option⁴

If the stock price doesn't change there will still be a change in option value as time progresses towards expiration. The *theta* of an option, Θ , is the rate of change of the option price with time. Note that theta for a long option will always be negative, since options become less valuable as they approach expiry. It is possible for an option (or portfolio of options) to have positive theta if it is held short (or the portfolio is dominated by short options). In this case the passage of time will result in profits. The issue of time decay is generally most significant for options that are close to expiry and at-the-money.

In practice, the volatility of the underlying is not known with certainty. Not only is it very difficult to measure at any time, it is even harder to predict what it will do in the future. Suppose that we put a volatility of 20% into an option pricing formula, how sensitive is the price to that number? How much would the price change if we changed the volatility, say to 21%? The sensitivity of the option price to the volatility is known as the *vega*. Long (short) options have positive (negative) vega, meaning that they will profit (incur losses from) an increase in volatility. Generally the longer the term of the option, the greater its sensitivity to changes in the volatility.⁵

Finally we mention *rho*, ρ , which is the sensitivity of the option value to the interest rate. Of course there are many more Greeks, but those we have mentioned are the main ones.

⁴ For the code to produce this surface plot see <http://www.mathworks.com/access/helpdesk/help/toolbox/finance/samples8.html>

⁵ As can be seen from Table I.A.8.2, the Black-Scholes vega behaves similarly to the Black-Scholes gamma – except that it *decreases* as the option approaches expiry – both take the shape of a normal density curve.

In Chapter II.C, when we explain what a *Taylor Expansion* is, we shall show how the change in value of an option, or a portfolio of options, can be approximated using the Greeks. Specifically, Example II.C.5.3 gives the ‘delta-gamma-vega’ approximation to the value of an option (or a portfolio of options). Then in Section II.D.2.3 we explain how the Greeks are used for hedging options portfolios.

When an option value is obtained using the Black-Scholes formula we can derive some simple formulae for the associated values of the Greeks. Here is a table of all these formulae for the main Greeks that were discussed above. Of course, Table I.A.8.2 is only for European options that are priced using the Black-Scholes model. When a different option pricing model is used, the sensitivities are not than same as those listed here. The functions N , d_1 and d_2 are as above:

Table I.A.8.2: Black-Scholes Greeks for European Calls and Puts

Greek	Call	Put
Value, V	$SN(d_1) - Ke^{-r(T-t)}N(d_2)$	$-SN(-d_1) + Ke^{-r(T-t)}N(-d_2)$
Delta, $\Delta = \frac{\partial V}{\partial S}$	$N(d_1)$	$N(d_1) - 1$
Gamma, $\Gamma = \frac{\partial^2 V}{\partial S^2}$	$\frac{N'(d_1)}{\sigma S \sqrt{T-t}}$	$\frac{N'(d_1)}{\sigma S \sqrt{T-t}}$
Theta, $\Theta = \frac{\partial V}{\partial t}$	$-\frac{\sigma SN'(d_1)}{2\sqrt{T-t}} - rEe^{-r(T-t)}N(d_2)$	$-\frac{\sigma SN'(-d_1)}{2\sqrt{T-t}} + rEe^{-r(T-t)}N(-d_2)$
Vega, $\frac{\partial V}{\partial \sigma}$	$S\sqrt{T-t} N'(d_1)$	$S\sqrt{T-t} N'(d_1)$
Rho, $\rho = \frac{\partial V}{\partial r}$	$E(T-t)e^{-r(T-t)}N(d_2)$	$-E(T-t)e^{-r(T-t)}N(-d_2)$

I.A.8.9 Implied Volatility

The Black-Scholes formula for a call option takes as input the expiry, the strike, the underlying and the interest rate, together with the volatility to output the price. All but the volatility are easily measured. How do we know what volatility to put into the formula?

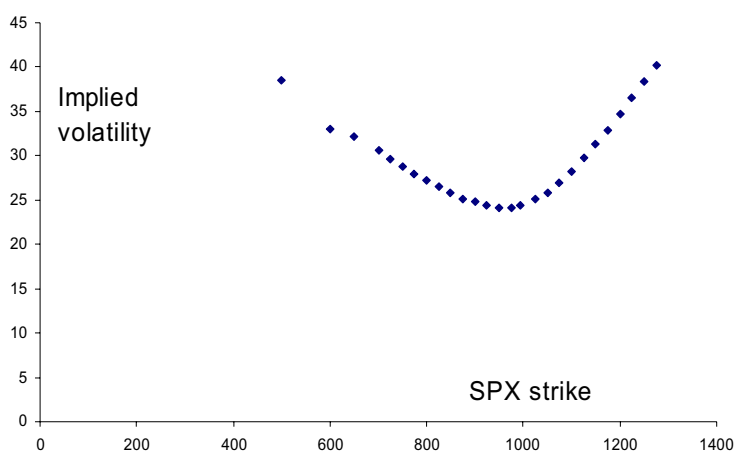
A trader can see on her screen that a certain call option with four months until expiry and a strike of 100 is trading at 6.51 with the underlying at 101.5 and a short-term interest rate of 4%. Can we use this information in some way?

Turn the relationship between volatility and an option price on its head. If we can see the price at which the option is trading, we can ask, ‘What volatility must I use to get the correct market

price?'. This is called the *implied volatility*. The implied volatility is the volatility of the underlying, which, when substituted into the Black–Scholes formula, gives a theoretical price equal to the market price. In a sense it is the market's view of volatility over the life of the option.

In practice, if we calculate the implied volatility for many different strikes and expiries on the same underlying, we find that the volatility is not constant. A typical result is that of Figure I.A.8.12, which shows the implied volatilities versus strike. The implied volatilities for the calls and puts should be identical, because of put–call parity.

Figure I.A.8.12



This shape is commonly referred to as the smile; the slope is known as the skew. The shape tends to persist with time, with certain shapes being characteristic of certain markets.

This shape can be interpreted in many ways: (1) the Black–Scholes model is right and the market is wrong; (2) the Black–Scholes model is wrong and the market ‘knows best’ about what prices should be; (3) the model is wrong and the market does not really know best. Clearly (3) is correct. Since no one really knows what volatility is, or will be, the market does what it is supposed to, that is, put buyers and sellers together. The result is that option prices respond to supply and demand just like Ferraris or groceries.

In the equity markets many people want to buy cheap out-of-the-money puts for protection in the event of a stock fall. The demand pressure causes put prices to rise, and so we get high implied volatilities for low strikes. There are also many people selling covered calls to earn a little bit extra. So supply pressure causes out-of-the-money calls to be relatively cheap. The end result for equities and indices is a negative skew, a downward-sloping implied volatility versus strike curve.

I.A.8.10 Intrinsic versus Time Value

An option's value can be interpreted as being made up of two parts, intrinsic value and time value.

Intrinsic value is what the payoff would be if the option could be exercised now, with the stock at its current price. If the option is out of the money then the intrinsic value is zero. *Time value* is then the difference between the option's value and its intrinsic value. Although a vanilla option's intrinsic value can never be negative, its time value can be positive or negative. This is because of the different effects that time has on a contract. Time generally increases value due to random fluctuations in the stock, but decreases value because we have to wait longer to get any payoff.

References

Black, F, and Scholes, M (1973) The pricing of options and corporate liabilities, *Journal of Political Economy*, **81**, 637–59.

Cox, J, and Rubinstein, M (1985) *Options Markets* (Englewood Cliffs, New Jersey: Prentice Hall).

Haug, EG (1998) *The Complete Guide to Option Pricing Formulas* (New York: McGraw-Hill).

Merton, RC (1973) Theory of rational option pricing, *Bell Journal of Economics and Management Science*, **4**, 141–83.

Kassouf, S, and Thorp, EO (1967) *Beat the Market* (New York: Random House).

Tudball, D (2001) In for the count, *Wilmott magazine*, September.

Wilmott, P (2000) *Paul Wilmott on Quantitative Finance* (Chichester: Wiley).

I.B.1 General Characteristics of Bonds

Lionel Martellini and Philippe Priaulet¹

Fixed-income markets are populated with a vast range of instruments. In this chapter we provide a typology of the simplest of these instruments, namely bonds and money-market instruments, and describe their general characteristics. The pricing of bonds and measures of price sensitivity are discussed in Chapter I.B.2.

I.B.1.1 Definition of a Bullet Bond

A debt security (or a bond) is a financial claim by which the issuer (or the borrower) is committed to paying back to the bondholder (or the lender) the cash amount borrowed (called the principal), plus periodic interest calculated on this amount during a given period of time. It can have either a bullet or a non-bullet structure. A bullet bond is a fixed-coupon bond without any embedded option, delivering its coupons on periodic dates and the principal on the maturity date.

As an example, a US Treasury bond with coupon 3.5%, maturity date 15 November 2006 and a nominal issued amount of \$18.8 billion pays a semi-annual interest of \$329 million (\$18.8 billion times 3.5%/2) every six months until 15 November 2006 inclusive, as well as \$18.8 billion on the maturity date. Another example would be a Euro Treasury bond with coupon 4%, maturity date 4 July 2009 and a nominal issued amount of €1 billion, which pays an annual interest of €440 million (€1 billion times 4%) every year until 4 July 2009 inclusive, as well as €1 billion on the maturity date.

The purpose of a bond issuer (the Treasury Department, a government entity or a corporation) is to finance its budget or investment projects (construction of roads, schools, development of new products, new plants) at an interest rate that is expected to be lower than the return rate of investment (at least for the private sector). Through the issuance of bonds, it has a direct access to the market, and so avoids borrowing from investment banks at higher interest rates. This process of financial disintermediation has gathered momentum in recent times. One point to underscore is that the bondholder has the status of a creditor, unlike the equity holder who has the status of an owner of the issuing corporation. This is, by the way, the reason why a bond is, generally speaking, less risky than an equity.

¹ Lionel Martellini is a Professor of Finance at EDHEC Graduate School of Business, and the Scientific Director of EDHEC Risk and Asset Management Research Center (www.edhec-risk.com). Philippe Priaulet is a Fixed-Income Strategist, in charge of derivatives strategies for HSBC, and also an Associate Professor in the Department of Mathematics of the University of Evry Val d'Essonne and a lecturer at ENSAE.

*Reproduced with kind permission of John Wiley & Sons Ltd from *Fixed-Income Securities: Valuation, Risk Management and Portfolio Strategies*, 2003.

I.B.1.2 Terminology and Convention

A bond issue is characterised by many components, as follows:

- *The issuer's name.* For example, Bundesrepublik Deutschland for a Treasury bond issued in Germany.
- *The issuer's type.* This is mainly the sector it belongs to, for example the oil sector if Total Fina Elf is the bond issuer.
- *The market in which the bond is issued.* This can be the US domestic market, the euro zone domestic market, the domestic market of any country, the Eurodollar market which corresponds to bonds denominated in US dollars that are issued in European countries, or any other market.
- *The issuer's country of origin.*
- *The bond's currency denomination.* An example is US dollars for a US Treasury bond.
- *The type of collateral.* This is the type of asset pledged against possible default. The collateral type can be a mortgage on property, an automobile loan, a government guarantee, and so forth.
- *The maturity date.* This is the date on which the principal amount is due.
- *The maturity type.* Some issues are callable prior to term at predetermined prices and times. That is, the issuer may elect to buy back the debt issue.
- *The coupon type.* It can be fixed, floating, multi-coupon (a mix of fixed and floating or different fixed). For example, a step-up coupon bond is a kind of multi-coupon bond with a coupon that increases at predetermined intervals.
- *The coupon rate.* This is expressed as a percentage of the principal amount when the coupon is fixed.
- *The coupon frequency.* The coupon frequency for Treasury bonds is semi-annual in the USA, the UK and Japan, and annual in the euro zone, except for Italy where it is semi-annual.
- *The interest rate type.* Interest rates can be *nominal* or *real*. When interest rates are nominal rather than real, this means that the effects of inflation have not been removed. For example, a nominal rate of return of 8%, with a 3% rate of inflation, implies a real rate of return of 5%.
- *The day count type.* The most common types are actual/actual, actual/365, actual/360 and 30/360. Actual/actual (actual/365, actual/360) means that the accrued interest between two

given dates is calculated using the exact number of calendar days between the two dates divided by the exact number of calendar days of the relevant year (365, 360). 30/360 means that the number of calendar days between the two dates is computed assuming that each month counts as 30 days. For example, using the 30/360 day count basis there are 84 days ($2 \times 30 + 24$) from 1 January 2001 to 25 March 2001 and 335 days ($11 \times 30 + 5$) from 1 January 2001 to 6 December 2001. Using the actual/actual or actual/365 day count basis, there are 83 days from 1 January 2001 to 25 March 2001 and 339 days from 1 January 2001 to 6 December 2001. Using the actual/actual day count basis, the period converted in years from 1 August 1999 to 3 September 2001 is $\{152/365\} + 1 + \{246/365\} = 2.0904$. Using the actual/365 day count basis, the period converted in years from 1 August 1999 to 3 September 2001 is $764/365 = 2.0931$. Using the actual/360 day count basis, the period converted in years from 1 August 1999 to 3 September 2001 is $764/360 = 2.1222$. Using the 30/360 day count basis, the period converted in years from 1 August 1999 to 3 September 2001 is $752/360 = 2.0888$.

- *The method used for the calculation of the bond yield.* There are different ways of calculating the yield to maturity, depending on the day count convention and style of compounding. In Example I.B.1.4 the Bloomberg Yield Analysis screen displays a variety of yield calculations. The most common of these, the ‘street convention’, uses simple interest in the first period and compounded interest thereafter. It uses the actual/actual day count.
- *The method used for the calculation of the bond price.* In most (but not all) markets the bond price is quoted as a ‘clean’ price, that is, excluding accrued interest. When a bond is bought/sold, however, the invoice price should include accrued interest (see Section I.B.1.3.1).
- *The announcement date.* This is the date on which the bond is announced and offered to the public.
- *The interest accrual date.* This is the date when interest begins to accrue.
- *The settlement date.* This is the date on which payment is due in exchange for the bond. It is generally equal to the trade date plus a number of working days. For example, in Japan, the settlement date for Treasury bonds and T-bills is equal to the trade date plus three working days. On the other hand, in the US, the settlement date for Treasury bonds and T-bills is equal to the trade date plus one working day. In the UK, the settlement date for Treasury bonds and T-bills is equal to the trade date plus one and two working days, respectively. In the euro zone, the settlement date for Treasury bonds is equal to the trade date plus three working days and for T-bills one, two or three working days depending on the country under consideration.

- *The first coupon date.* This is the date of the first interest payment.
- *The issuance price.* This is the price paid at issuance expressed as a percentage of par value.
- *The spread at issuance.* This is the difference between the yield on the bond and the yield on a benchmark Treasury bond (expressed in basis points).
- *The identifying code.* The most popular ones are the ISIN (International Securities Identification Number) and CUSIP (Committee on Uniform Securities Identification Procedures) numbers.
- *The rating.* The task of rating agencies is to assess the default probability of corporations through what is known as rating. A rating is a ranking of a bond's quality, based on criteria such as the issuer's reputation, management, balance sheet and record in paying interest and principal. The two major ones are Moody's and Standard and Poor's (S&P). Their rating scales are listed in Table I.B.1.1. We refer the reader to Martellini *et al.* (2003) for more details.

Table I.B.1.1: Moody's and S&P's rating scales

Investment grade (high creditworthiness)		
Moody's	S&P	Definition
Aaa	AAA	Gilt-edged, best quality, extremely strong creditworthiness
Aa1 Aa2 Aa3	AA+ AA AA–	Very high grade, high quality, very strong creditworthiness
A1 A2 A3	A+ A A–	Upper medium grade, strong creditworthiness
Baa1 Baa2 Baa3	BBB+ BBB BBB–	Lower medium grade, adequate creditworthiness
Speculative grade (low creditworthiness)		
Moody's	S&P	Definition
Ba1 Ba2 Ba3	BB+ BB BB–	Low grade, speculative, vulnerable to non-payment
B1 B2 B3	B+ B B–	Highly speculative, more vulnerable to non-payment
Caa	CCC+ CCC CCC–	Substantial risk, in poor standing, currently vulnerable to non-payment
Ca C	CC C	May be in default, extremely speculative, currently highly vulnerable to non-payment
	D	Even more speculative Default

The modifiers 1, 2, 3 or +, – account for relative standing within the major rating categories.

- *The total issued amount.* This is given in thousands of the issuance currency on Bloomberg.
- *The outstanding amount.* This is the amount of the issue still outstanding, which appears in thousands of the issuance currency on Bloomberg.
- *The minimum amount and minimum increment that can be purchased.* The minimum increment is the smallest additional amount of a security that can be bought above the minimum amount.
- *The par amount or nominal amount or principal amount.* This is the face value of the bond. Note that the nominal amount is used to calculate the coupon bond. For example, consider a bond with a fixed 5% coupon rate and a \$1000 nominal amount. The coupon is equal to $5\% \times \$1000 = \50 .
- *The redemption value.* Expressed as a percentage of the nominal amount, this is the price at which the bond is redeemed on the maturity date. In most cases, the redemption value is equal to 100% of the bond nominal amount.

We now give some examples of a Bloomberg bond description screen (DES function), for Treasury and corporate bonds.

Figure I.B.1.1: Bloomberg screen for US Treasury bond

SECURITY INFORMATION		ISSUER INFO	REDEMPTION INFO
CPN FREQ	2	NAME	US TREASURY N/B
CPN TYPE	FIXED	TYPE	US GOVT NATIONAL
MTY/REFUND TYP	NORMAL	IDENTIFICATION #'s	
CALC TYP (1)STREET CONVENTION		CUSIP	9128277F3
DAY COUNT (1)ACT/ACT		MLNUM	H2665
MARKET ISS	US GOVT	SEDOL 1	2817479
COUNTRY/CURR	USA/ DOL	WERTPAP	777622
SECURITY TYPE	USN	ISIN	US9128277F31
AMT ISSUED	18801(MM)	EURO COM	013883777
AMT OUTSTAND	18801(MM)	ISSUANCE INFO	
MIN PIECE	1000	ISSUE DATE	11/15/01
		INT ACCRUES	11/15/01
		1ST CPN DT	5/15/02
		PRC @ ISSUE	99.469
		PRICE FORMAT	
		32-nds	96-5
		Decimal	96.15625000
		Repurch Pgm	
TENDERS ACCEPTED: \$16000MM.			
<small> Australia 61 2 9777 8600 Brazil 5511 3048 4500 Europe 44 20 7330 7500 Germany 49 69 92041210 Hong Kong 852 2977 6000 Japan 81 3 3201 8900 Singapore 65 212 1000 U.S. 1 212 318 2000 Copyright 2001 Bloomberg L.P. 1356-711-0 10-Dec-01 12:03:00 </small>			

Example I.B.1.1

The T-bond, with coupon rate 3.5% and maturity date 15 November 2006 (see Figure I.B.1.1), bears a semi-annual coupon with an actual/actual day count basis. The issued amount was equal to \$18.8 billion; the outstanding amount is still \$18.8 billion. The minimum amount that can be purchased is equal to \$1000. The T-bond was issued on 15 November 2001 in the US market, and interest began to accrue from this date on. The issue price was 99.469. The first coupon date was 15 May 2002, that is, 6 months after the interest accrual date (semi-annual coupon). This bond has an AAA rating.

Example I.B.1.2

In comparison with the previous bond, the Elf bond has a Aa2 Moody's rating (see Figure I.B.1.2). It belongs to the oil sector. The issued amount was €1 billion and the minimum purchasable amount is €1000. The issue price was 98.666. It delivers an annual fixed 4.5% coupon rate. Its maturity is 23 March 2009. Its spread at issuance amounted to 39 basis points over the French T-bond (OAT) with coupon 4% and maturity date 25 April 2009.

Figure I.B.1.2: Bloomberg screen for Elf bond

SECURITY DESCRIPTION		DL19 Corp	DES
ELF AQUITAINE FFPF 4 1/2 03/09 95.7440/96.2440		Page 1 / 1 (5.21/5.13) BGN @12/07	
ISSUER INFORMATION		IDENTIFIERS	
Name	ELF AQUITAINE	Common	009552197
Type	Oil Comp-Integrated	ISIN	XS0095521976
Market of Issue	EURO-ZONE	French	049452
SECURITY INFORMATION		RATINGS	
Country	FR	Currency	EUR
Collateral Type	SR UNSUB	Moody's	Aa2
Calc Typ	(962)STREET CONVENTION	S&P	NR
Maturity	3/23/2009 Series	Composite	AA2
NORMAL		ISSUE SIZE	
Coupon	4 1/2 FIXED	Amt Issued	EUR 1,000,000 (M)
ANNUAL	ACT/ACT	Amt Outstanding	EUR 1,000,000 (M)
Announcement Dt	3/ 4/99	Min Piece/Increment	1,000.00/ 1,000.00
Int. Accrual Dt	3/23/99	Par Amount	1,000.00
1st Settle Date	3/23/99	BOOK RUNNER/EXCHANGE	
1st Coupon Date	3/23/00	BNP, GS	
Iss Pr	98.6660 Reoffer 98.666	LONDON	
SPR @ FPR	39.0 vs FRTR 4 04/09	65) Old DES	
NO PROSPECTUS		66) Send as Attachment	
UNSEC'D, SEASONED EFF 5/02/99,			
<small>Australia 61 2 9777 8600 Brazil 5511 3048 4500 Europe 44 20 7330 7500 Germany 49 69 92041210 Hong Kong 852 2977 6000 Japan 81 3 3201 8900 Singapore 65 212 1000 U.S. 1 212 318 2000 Copyright 2001 Bloomberg L.P. 1356-711-0 10-Dec-01 12:06:00</small>			

I.B.1.3 Market Quotes

Bond securities are usually quoted in terms of price (percentage price of the nominal amount), yield or spread over an underlying benchmark bond.

I.B.1.3.1 Bond Quoted Price

The quoted price (or market price) of a bond is usually its *clean price*, that is, its *gross price* minus the *accrued interest*. Note that these two prices are usually given as percentages of the nominal amount.² When an investor purchases a bond, he is entitled to receive all the future cash flows of this bond, until he no longer owns it. If he buys the bond between two coupon payment dates, he logically must buy it at a price reflecting the fraction of the next coupon that the seller of the bond is entitled to receive for having held it until the sale. This price is called the *gross price* (or *dirty price* or *full price*). It is computed as the sum of the clean price and the portion of the coupon that is due to the seller of the bond. This portion is called the *accrued interest*. Note that the accrued interest is computed from the last coupon date to the settlement date. The settlement date is equal to the transaction date plus n working days, n depending on the bond market type. For instance, n is equal to 1 for US Treasury bonds.

Example I.B.1.3

On 10 December 2001 an investor buys a given amount of the US Treasury bond with coupon 3.5% and maturity 15 November 2006. The current clean price is 96.15625. Hence the market value of \$1 million face value of this bond is equal to $96.15625\% \times \$1 \text{ million} = \$961,562.50$. The accrued interest period is equal to 26 days. Indeed, this is the number of calendar days between the settlement date (11 December 2001) and the last coupon payment date (15 November 2001). Hence the accrued interest is equal to the last coupon payment (1.75, because the coupon frequency is semi-annual) times 26 divided by the number of calendar days between the next coupon payment date (15 May 2002) and the last coupon payment date (15 November 2001). In this case, the accrued interest is equal to $1.75\% \times 26/181 = 0.25138\%$. The gross price is then 96.40763. The investor will pay \$964,076.30 ($96.40763\% \times \1 million) to buy this bond.

Note that the *clean price* of a bond is equal to the gross price on each coupon payment date and that US bond prices are commonly quoted in multiples of $1/32$. For example, a bond quoting a price of 98-28 actually has a price of $98 + 28/32 = 98.875$. In most other markets decimal quotes are the norm.

² When the bond price is given as a dollar (euro, sterling, etc.) amount, it is the nominal amount of the bond multiplied by the price as a percentage of the nominal amount.

I.B.1.3.2 Bond Quoted Yield

The quoted yield of a bond is the discount yield that equates its gross price times its nominal amount to the sum of its discounted cash flows.

Example I.B.1.4 Bond yield quotes on Bloomberg

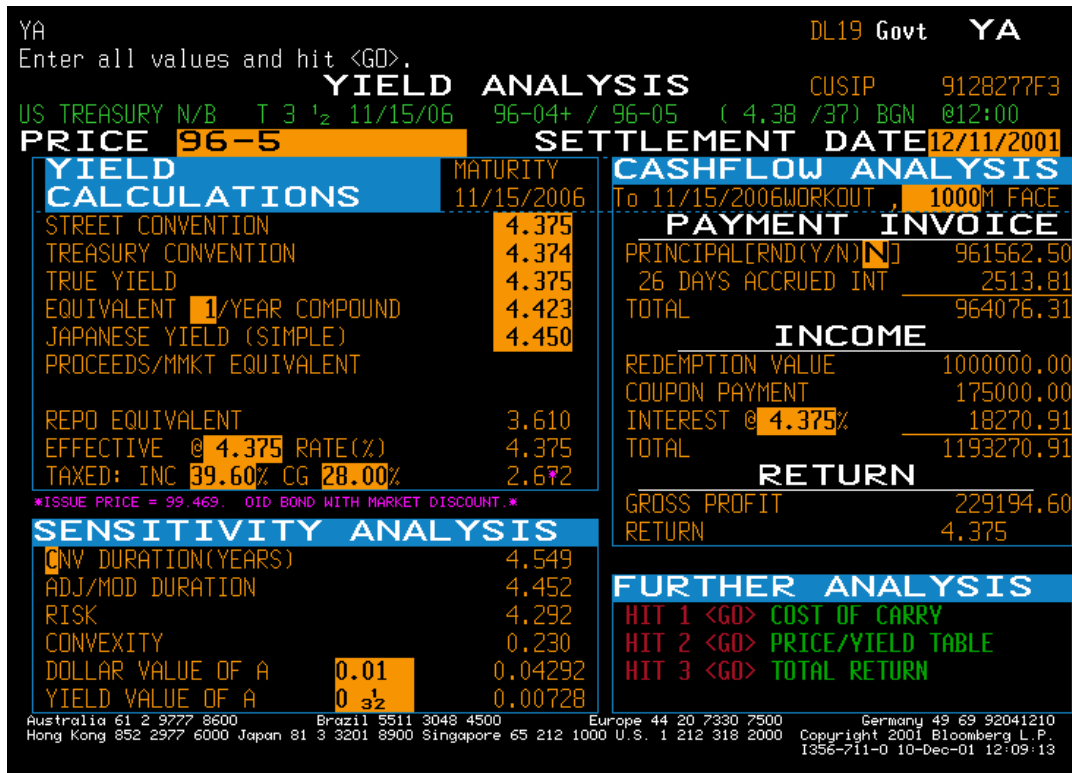
In the previous example, the cash-flow schedule of the bond with \$1 million face value is as shown in Table I.B.1.2.

Table I.B.1.2

<i>Date</i>	<i>Cash Flow</i>	<i>Date</i>	<i>Cash Flow</i>
15 May 2002	17,500	15 Nov 2004	17,500
15 Nov 2002	17,500	15 May 2005	17,500
15 May 2003	17,500	15 Nov 2005	17,500
15 Nov 2003	17,500	15 May 2006	17,500
15 May 2004	17,500	15 Nov 2006	1,017,500

The Bloomberg Yield Analysis screen (YA function) associated with this bond is shown in Figure I.B.1.3. Its quoted yield is equal to 4.375% using the standard street convention (as explained in Section I.B.1.2). Notice that the yields calculated using the Treasury convention and true yield are similar; the latter adjusts for non-business days by moving the coupon date to the next valid business date when necessary. The equivalent one-year compound yield of this bond is equal to 4.423%. This method is similar to the street convention, but converts the yield from semi-annual to annual compounding. Finally, the Japanese yield is a simple yield calculation, whereby the annualised cash flow is expressed as a percentage of the original clean price.

Figure I.B.1.3: Bloomberg screen for Japanese bond



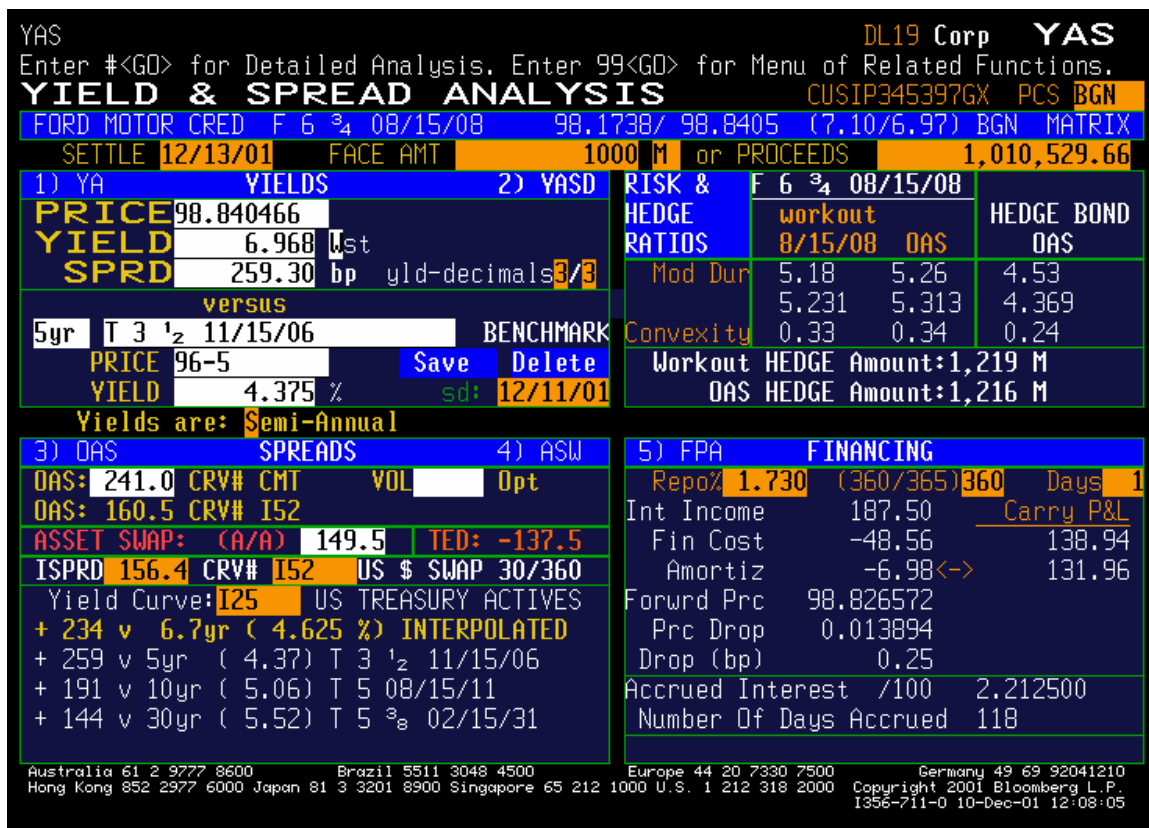
I.B.1.3.3 Bond Quoted Spread

Corporate bonds are usually quoted in price or in spread over a given benchmark bond rather than in yield. To recover the corresponding yield, you simply have to add this spread to the yield of the underlying benchmark bond.

Example I.B.1.5

The Bloomberg screen in Figure I.B.1.4 gives an example of a bond yield and spread analysis as can be seen on a Bloomberg screen. The bond was issued by Ford Motor. It bears a spread of 156.4 basis points (see ISPRD function) over the interpolated US dollar swap yield, whereas it bears a spread of 234 basis points over the interpolated US Treasury benchmark bond yield. Furthermore, its spread over the US Treasury benchmark bond with the nearest maturity amounts to 259.3 basis points (SPRD function). It is 191 and 144 basis points over the 10-year Treasury benchmark bond and the 30-year Treasury benchmark bond, respectively.

Figure I.B.1.4: Bloomberg screen for yield and spread analysis



Note that this last way of quoting spreads relative to government bonds is fairly common in bond markets. Indeed, in this case the underlying government bond is clearly identified, while using an interpolation on the government bond curve may lead to different results, depending on the two bonds as well as the kind of interpolation considered.

I.B.1.3.4 Liquidity Spreads

Investors generally prefer to invest in more liquid securities which offer opportunities for trading in larger volumes with lower bid–offer spreads. A liquid Treasury note/bond is known as an ‘on-the-run’ issue, whereas its less liquid counterpart is an ‘off-the-run’ issue. The liquidity of an issue is primarily a function of the time since issue, with seasoned issues tending to be less liquid. More recent issues are more liquid as dealers have more information about who owns the securities and in what amounts. Consequently, trading volumes in these ‘on-the-run’ securities are higher and bid–offer spreads can be as little as 1/32. Yields of ‘off-the-run’ bonds exceed those of ‘on-the-run’ bonds even though their duration and credit rating may be identical. This additional yield may be thought of as a risk premium to compensate the investor for liquidity risk. In a liquidity crisis, such as that which occurred in 1998, liquid securities are even more highly prized and the liquidity spread widens.

I.B.1.3.5 The Bid–Ask Spread

To finish, it is worth mentioning that every traded bond has a bid as well as an ask quoted price. The bid price is the price at which an investor can sell a bond, whereas the ask price is the price at which he can buy it. The ask price is of course higher than the bid price, which means that the ask yield is lower than the bid yield. The difference between the two yields is known as the bid–ask spread. It is a kind of transaction cost. It is very small for liquid bonds like US or Euro Treasury bonds, and is large for fairly illiquid bonds. The bond's mid price is simply the average of its bid and ask prices. The same holds for the mid yield.

I.B.1.4 Non-bullet Bonds

In this section we discuss strips, floating-rate notes and inflation-indexed bonds.

I.B.1.4.1 Strips

Strips (separate trading of registered interest and principal) are zero-coupon bonds mainly issued by government bonds of the G7 countries. The strips programme was created in 1985 by the US Treasury Department in response to investment banks who in the early 1980s had been buying long-term Treasury bonds and then issuing their own zero-coupon bonds to the public, collateralised by the payments on the underlying Treasury bonds. These so called trademark zeros were a success, but because of the higher liquidity of strips they were dominated by them. The only cash flow distributed by strips is the principal on the maturity date.

Example I.B.1.6

An investor buys for \$20,000 the Treasury strip bond with maturity 15 May 2030, and nominal amount \$100,000. As a bondholder, he is entitled to receive back \$100,000 on 15 May 2030, if he has of course not sold the bond meanwhile.

Such a bond that yields no coupon interest over the investment period may seem rather peculiar and unattractive. In fact, it effectively bears interest on maturity as it is bought at a price that is lower than its maturity price. The investors who buy these bonds are usually long-term investors such as pension funds and insurance companies and have at least one main purpose, which is to secure a return over their long-term investment horizon. To understand this point, consider an investor who is supposed to guarantee 6% per annum over 20 years on their liabilities. If the investor buys and holds a strip with a maturity equal to his/her investment horizon, that is 20 years, and a yield to maturity of 6%, he/she meets his objective perfectly because he/she knows today the return per annum on that bond, which is 6%. In contrast, coupon-bearing bonds do not allow him/her to do so – first, because they bear an interest reinvestment risk, and second, because their duration rarely if ever reaches 20 years.

There exist two types of strips, *coupon strips* and *principal strips*. Coupon strips and principal strips are built through stripping the coupons and the principal of a coupon-bearing bond, respectively. The main candidates for stripping are government bonds (Treasury bonds and government agency bonds). Strips are not as liquid as coupon-bearing bonds. Their bid–ask spread is usually higher.

I.B.1.4.2 Floating-Rate Notes

Floating-rate notes (FRNs) are bond securities that bear floating coupon rates. Actually, this generic denomination encompasses two categories of bonds:

- floating-rate bonds;
- variable-rate bonds or adjustable-rate bonds.

The former category denotes bonds whose coupon rates are indexed to a short-term reference with a maturity less than one year, like the three-month Libor, whereas the latter designates bonds whose coupon rates are indexed to a longer-term reference with a maturity greater than one year, like the 10-year constant maturity Treasury (CMT) bond yield (see Example I.B.1.9). The coupons of floating-rate bonds are reset more than once a year. This is not necessarily the case for variable-rate bonds, which may have a reset frequency exceeding one year. Usually, the reset frequency is equal to the coupon payment frequency.

Furthermore, FRNs differ from each other as regards the nature of the coupon rate indexation. Coupon rates can be determined in three ways:

- as the product of the last reference index value and a multiplicative margin;
- as the sum of the last reference index value and an additive margin;
- as a mix of the two previous methods.

Note that when the sign of the multiplicative margin is negative, the bond is called an inverse floater. The coupon rate moves in the opposite direction to the reference index. So as to prevent it from becoming negative, a floor is determined that is usually equal to zero. Such bonds have become fairly popular in the context of decreasing interest rates. Let us now present some examples.

Example I.B.1.7

An investor buying a floating-rate bond whose coupon rate is equal to three-month Libor + 20 bp is entitled to receive, every period determined in the contract (usually every three months), a coupon payment inversely proportional to its annual frequency and principal on maturity date.

The coupon rate will be reset every three months in order to reflect the new level of the three-month Libor rate.

Example I.B.1.8

An investor buying an inverse floater whose coupon rate is equal to $16\% - 2x$, where x is the two-year T-bond yield, is entitled to receive, every period determined in the contract (usually every year), a coupon payment inversely proportional to its annual frequency and principal on the maturity date. The coupon rate will be reset every two years in order to reflect the new level of the two-year bond yield.

Example I.B.1.9: The French 10-year CMT bond description on Bloomberg

The French 10-year CMT bond with maturity date 25 October 2006 bears a quarterly floating coupon which is indexed on TEC 10 (see Figure I.B.1.5). TEC 10 is a French 10-year CMT reference. It is determined on a daily basis as the 10-year interpolated yield between two active Treasury bond yields with nearest maturity. The coupon rate is equal to TEC10 minus 100 bp and entitles the bondholder to receive every quarter on 25 January, 25 April, 25 July and 25 September a coupon payment equal to $(1 + \text{TEC10} - 100 \text{ bp})^{1/4} - 1$, and the principal, on the maturity date 25 October 2006. Coupon rates are reset every quarter on an actual/actual day count basis. For example, the coupon paid on 25 April is determined using the TEC 10 index five working days before 25 January. The issued amount is equal to €11.888 billion, like the outstanding amount. The minimum amount that can be purchased is €1. The bond was issued in the euro zone. It has a AAA rating. Its issue price was 101.55. The bid and ask prices on 13 December 2001 were 99.504 and 99.5665, respectively.

Figure I.B.1.5: French 10-year CMT bond description on Bloomberg

ISSUER INFORMATION		IDENTIFIERS		1) Euro Redenomination 2) Additional Sec Info 3) Floating Rates 4) Identifiers 5) Ratings 6) Custom Notes 7) Issuer Information 8) ALLQ 9) Pricing Sources 10) Related Securities 65) Old DES 66) Send as Attachment
FRANCE O.A.T. FRTR Float 10/06 99,5040/99,5665		Common 008960194		
Name FRANCE (GOVT OF)		ISIN FR0000570541		
Type Sovereign		French 057054		
Market of Issue EURO-ZONE				
SECURITY INFORMATION		RATINGS		
Country FR Currency EUR		Moody's Aaa		
Collateral Type BONDS		S&P AAA		
Calc Typ(624)TEC10:FFR VAR NOTE		Composite AAA		
Maturity 10/25/2006 Series TC10		ISSUE SIZE		
NORMAL		Amt Issued		
Coupon 3.76 FLOATING QUARTLY		EUR 11,887,669 (M)		
TEC10 -100 ACT/ACT		Amt Outstanding		
Announcement Dt 4/12/96		EUR 11,887,669 (M)		
Int. Accrual Dt 4/25/96		Min Piece/Increment		
1st Settle Date 4/25/96		1.00/ 1.00		
1st Coupon Date 7/25/96		Par Amount 1.00		
Iss Pr 101.5500		BOOK RUNNER/EXCHANGE		
NO PROSPECTUS		BNP/CDC		
		EURONEXT-PARIS		
CPN RATE=TEC10 -100BP. ORIG F#18BLN ISS'D 4/25/96. ADD'L F#8.155BLN ISS'D 5/24/96, F#7.146BLN 6/25/96, F#6.124BLN 7/96, F#8.916 9/96, F#8.264 10/96, F#8.32BLN				

When buying an FRN an investor is typically hedged against parallel shifts of the interest-rate curve, because the coupons of the bond reflect the new level of market interest rates on each reset date. So, FRNs usually outperform fixed-rate bonds with the same maturity when interest rates shift upwards, and underperform them when interest rates shift downwards. Regarding inverse floaters, the situation is more complex due to the way they are structured. A decrease in interest rates will not necessarily result in the price appreciation of inverse floaters, despite the increase in the coupon rate. Their performance depends in fact on the evolution of the shape of the interest-rate curve.

I.B.1.4.3 Inflation-Indexed Bonds

Inflation-indexed bonds deliver coupons and principal that are indexed to future inflation rates. They are structured so as to protect and increase an investor's purchasing power. They are mainly issued by governments signalling that they are willing to maintain a low inflation level. They are more developed in the UK where they represent over 20% of outstanding government bonds. In the USA they represented only 7% of the issued government debt in 1999. In France, there were only three Treasury inflation-indexed bonds (referred to as OATi) in December 2001.

The inflation rate between date t and date $t + 1$, denoted by $IR_{t,t+1}$ is defined as

$$IR_{t,t+1} = (CPI_{t+1}/CPI_t) - 1$$

where CPI_t is the consumer price index on date t .

The major characteristic of inflation-indexed bonds is that they deliver coupons and redemption values linked to the increase in the CPI. Let us examine the case of OATi bonds.

- The daily inflation reference on date t , denoted by DIR_t , is computed by using a linear interpolation of two CPIs as follows

$$DIR_t = CPI_{m-3} + \{(nt-1)/ND_m\} \{CPI_{m-2} - CPI_{m-3}\}$$

where CPI_m is the consumer price index of month m , ND_m is the number of days of month m and nt is the day of date t (e.g. for 26 April 2001 it is simply 26).

- The coupon payment of an OATi received on date t , denoted by C_t , is:

$$C_t = FV \times RC \times (DIR_t/DIR_{\text{initial}}).$$

where FV is the face value, RC is the real coupon and DIR_{initial} is the daily inflation reference on the initial date, which is a date varying with each OATi. For example, the initial date of the OATi maturing on 25 July 2029 is 25 July 1999.

- The redemption value of an OATi received on date T , denoted by RV_T , is obtained from the formula

$$RV_T = FV \times (DIR_T/DIR_{\text{initial}}).$$

- The accrued interest on an OATi on date t , denoted by AC_t , is

$$AC_t = C_t \times (\text{number of accrued days/exact number of days in the coupon period}).$$

Example I.B.1.10: A French Inflation-Indexed Treasury Bond Description on Bloomberg

The OATi with real coupon 3% and maturity date 07/25/2012, bears an annual coupon with an actual/actual day count basis (see Figure I.B.1.6). The issued amount is equal to €6.5 billion, like the outstanding amount. The minimum amount that can be purchased is €1. The first coupon date is 25 July 2002. This bond has a AAA rating. The bid and ask prices on 13 December 2001 were 99.09 and 99.22, respectively.

Figure I.B.1.6: French inflation-indexed Treasury bond description on Bloomberg

ISSUER INFORMATION		IDENTIFIERS		1) Additional Sec Info 2) Identifiers 3) Ratings 4) Sec. Specific News 5) Involved Parties 6) Custom Notes 7) Issuer Information 8) ALLQ 9) Pricing Sources 10) Related Securities 65) Old DES 66) Send as Attachment
Name	FRANCE (GOVT OF)	Common	013817669	
Type	Sovereign	ISIN	FR0000188013	
Market of Issue	EURO-ZONE	French	018801	
SECURITY INFORMATION		RATINGS		
Country	FR	Moody's	NA	
Currency	EUR	S&P	NA	
Collateral Type	DEBENTURES	Fitch	NA	
Calc Typ	(864)FRANCE I/L:STREET			
Maturity	7/25/2012 Series DATE	ISSUE SIZE		
NORMAL		Amt Issued		
Coupon	3 FIXED	EUR	6,500,000 (M)	
ANNUAL	ACT/ACT	Amt Outstanding		
Announcement Dt	10/23/01	EUR	6,500,000 (M)	
Int. Accrual Dt	7/25/01	Min Piece/Increment		
1st Settle Date	10/31/01		1.00/ 1.00	
1st Coupon Date	7/25/02	Par Amount	1.00	
Iss Pr	100.1730	BOOK RUNNER/EXCHANGE		
NO PROSPECTUS		BARCLY,DB,SG		
		EURONEXT-PARIS		
EURO-ZONE INFLATION INDEX LINKED BOND (INDEX LINKED TO CPXTEMU).				

An inflation-indexed bond can be used to hedge a portfolio, to diversify a portfolio or to optimise asset and liability management.

- When buying an inflation-indexed bond an investor is typically hedged against a rise in the inflation rate.
- This product is weakly correlated with other assets such as stocks, fixed-coupon bonds and cash, which makes it an efficient asset to diversify a portfolio.
- Insurance companies can use this product to hedge inflation risk between the time when a contingency occurs and the time when it is paid to the client; some pension funds guarantee their clients a performance indexed by the inflation rate, and so buy inflation-indexed bonds to reduce the mismatch between assets and liabilities.

I.B.1.5 Summary

Fixed-income markets are populated by a vast range of instruments. In this chapter, we provide a typology of the simplest of these instruments, namely bonds, and describe their general characteristics.

A bond is a financial claim by which the issuer, the borrower, is committed to paying back to the bondholder, the lender, the cash amount borrowed (called the principal), plus periodic interest

calculated on this amount over a given period of time. It can have either a bullet or a non-bullet structure. A bullet bond is a fixed-coupon bond with no embedded option, delivering its coupons on periodic dates and the principal on the maturity date. Non-bullet bonds such as strips, floating-rate notes and inflation-indexed bonds are also traded on bond markets. These bonds can be issued by government agencies, municipalities, or corporations. Bond quotes are usually expressed in terms of price, yield or spread over an underlying benchmark bond. The quoted price of a bond is usually its *clean price*, that is, its *gross price* minus the *accrued interest*. The quoted yield of a bond is the discount yield that equates its gross price times its nominal amount to the sum of its discounted cash flows. Corporate bonds are usually quoted in terms of price and spread over a given benchmark bond rather than yield; to recover the corresponding yield, you simply have to add this spread to the yield of the underlying benchmark bond.

Reference

Martellini, L, Priaulet, S, and Priaulet, P (2003) *Fixed-Income Securities: Valuation, Risk Management and Portfolio Strategies* (Chichester: Wiley).

I.B.2 The Analysis of Bonds

Moorad Choudhry¹

Bonds and shares form part of the *capital markets*. Shares are *equity capital* while bonds are *debt capital*. So bonds are a form of debt, much like a bank loan. Unlike bank loans, however, bonds can be *traded* in a market. A bond is a debt capital market instrument issued by a borrower, who is then required to repay to the lender/investor the amount borrowed plus interest, over a specified period of time. Bonds are also known as *fixed-income* instruments, or *fixed-interest* instruments in the sterling markets. Usually bonds are considered to be those debt securities with terms to maturity of over one year. Debt issued with a maturity of less than one year is considered to be *money market* debt.

There are many different types of bond that can be issued. The most common is the *conventional* (or *plain vanilla* or *bullet*) *bond*. This is a bond paying regular (annual or semi-annual) interest at a fixed rate over a fixed period to maturity or redemption, with the return of *principal* (the par or nominal value of the bond) on the maturity date. All other bonds will be variations on this. An introduction to the different types of bonds, and a summary of their general characteristics, is given in Chapter I.B.1.

A bond is a financial contract, in effect an IOU from the *issuer*, the person or body that has issued the bond. Unlike shares or equity capital, bonds carry no ownership privileges. An investor who has purchased a bond and thereby lent money to an institution will have no voice in the affairs of that institution and no vote at the annual general meeting. The bond remains an interest-bearing obligation of the issuer until it is repaid, which is usually the *maturity* date of the bond. The issuer can be anyone from a private individual to a sovereign government. A summary of the terminology and conventions used in bond issues is given in Section I.B.1.2.

There are a wide range of participants involved in the bond markets. We can group them broadly into borrowers and investors, plus the institutions and individuals who are part of the business of bond trading. Borrowers access the bond markets as part of their financing requirements; hence borrowers can include sovereign governments, local authorities, public-sector organisations and corporations. Virtually all businesses operate with a financing structure that is a mixture of debt and equity finance. The debt finance may well contain a form of bond finance, so it is easy to see what an important part of the global economy the bond markets are.

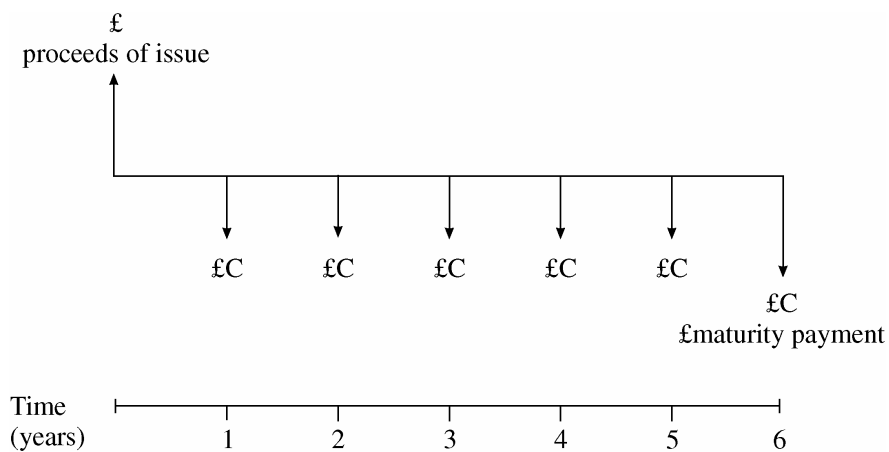
¹ Visiting Professor, Department of Economics, Finance and International Business, London Metropolitan University.

The different types of bond in the market reflect the different types of issuer and their respective requirements. Some bonds are safer investments than others. The advantage of bonds to an investor is that they represent a fixed source of current income, with an assurance of repayment of the loan on maturity. Bonds issued by developed-country governments are deemed to be guaranteed investments in that the final repayment is virtually certain. In the event of *default* of the issuing entity, bondholders rank above shareholders for compensation payments. There is lower risk associated with bonds than with shares as an investment, and therefore almost invariably a lower return in the long term.

I.B.2.1 Features of Bonds

A bond is a debt instrument, usually paying a fixed rate of interest over a fixed period of time. Therefore a bond is a collection of *cash flows*, and this is illustrated at Figure I.B.2.1. In our hypothetical example the bond, denominated in sterling, is a six-year issue that pays fixed interest payments of $C\%$ of the *nominal* value on an annual basis. In the sixth year there is a final interest payment and the loan proceeds represented by the bond are also paid back, known as the *maturity proceeds*. The amount raised by the bond issuer is a function of the price of the bond at issue, which we have labelled here as the issue proceeds.

Figure I.B.2.1: Cash flows associated with a six-year annual coupon bond



The upward-facing arrow in Figure I.B.2.1 represents the cash flow paid and the downward-facing arrows are the cash flows received by the bond investor. The cash flow diagram for a six-year bond that had a 5% fixed interest rate, known as a 5% *coupon*, would show interest payments of £5 per every £100 of bonds, with a final payment of £105 in the sixth year representing the last coupon payment and the redemption payment. Again, the amount of funds raised per £100 of bonds depends on the price of the bond on the day it is first issued, and we will look further

into this later. If our example bond paid its coupon on a semi-annual basis, the cash flows would be £2.50 every six months until the final redemption payment of £102.50.

Let us examine some of the key features of bonds.

I.B.2.1.1 Type of Issuer

A primary distinguishing feature of a bond is its issuer. The nature of the issuer will affect the way the bond is viewed in the market. There are four issuers of bonds: sovereign governments and their agencies; local government authorities; supranational bodies such as the World Bank; and corporations. Within the corporate bond market there is a wide range of issuers, each with differing abilities to satisfy their contractual obligations to investors. The largest bond markets are those of sovereign borrowers, the government bond markets. The United Kingdom government issues *gilts*. In the United States government bonds are known as *Treasury notes* and *Treasury bonds*, or simply *Treasuries*.

I.B.2.1.2 Term to Maturity

The *term to maturity* of a bond is the number of years after which the issuer will repay the obligation. During the term the issuer will also make periodic interest payments on the debt. The *maturity* of a bond refers to the date on which the debt will cease to exist, at which time the issuer will redeem the bond by paying the principal. The practice in the market is often to refer simply to a bond's 'term' or 'maturity'. The provisions under which a bond is issued may allow either the issuer or investor to alter a bond's term to maturity after a set notice period, and such bonds need to be analysed in a different way. The term to maturity is an important consideration in the make-up of a bond. It indicates the time period over which the bondholder can expect to receive the coupon payments and the number of years before the principal will be paid in full. The bond's *yield* also depends on the term to maturity. Finally, the price of a bond will fluctuate over its life as yields in the market change and as it approaches maturity. As we will discover later, the *volatility* of a bond's price is dependent on its maturity; assuming other factors constant, the longer a bond's maturity, the greater the price volatility resulting from a change in market yields.

I.B.2.1.3 Principal and Coupon Rate

The *principal* of a bond is the amount that the issuer agrees to repay the bondholder on the maturity date. This amount is also referred to as the *redemption value*, *maturity value*, *par value* or *face amount*, or simply *par*. The *coupon rate* or *nominal rate* is the interest rate that the issuer agrees to pay each year. The annual amount of the interest payment made is called the *coupon*. The coupon rate multiplied by the principal of the bond provides the cash amount of the coupon. For example, a bond with a 7% coupon rate and a principal of £1,000,000 will pay annual interest of £70,000. In

the United Kingdom, United States, Australia and Japan the usual practice is for the issuer to pay the coupon in two semi-annual instalments. For bonds issued in European markets and the Eurobond market coupon payments are made annually. On rare occasions one will encounter bonds that pay interest on a quarterly or monthly basis. All bonds make periodic interest payments, except for *zero-coupon bonds*. These bonds allow a holder to realise interest by being sold substantially below their principal value. The bonds are redeemed at par, with the interest amount then being the difference between the principal value and the price at which the bond was sold. We will explore zero-coupon bonds in greater detail later.

I.B.2.1.4 Currency

Bonds can be issued in virtually any currency. The largest volume of bonds in the global markets is denominated in US dollars; other major bond markets are denominated in euros, Japanese yen and sterling, and liquid markets also exist in Australian, New Zealand and Canadian dollars, Swiss francs and other major currencies. The currency of issue may impact on a bond's attractiveness and liquidity, which is why borrowers in developing countries often elect to issue in a currency other than their home currency, for example US dollars, as this will make it easier to place the bond with investors. If a bond is aimed solely at a country's domestic investors it is more likely that the borrower will issue in the home currency.

I.B.2.2 Non-conventional Bonds

The description of a bond in the previous section referred to a conventional or *plain vanilla* bond. There are many variations on vanilla bonds and we can introduce a few of them here.

I.B.2.2.1 Floating-Rate Notes

The bond market is often referred to as the *fixed-income* market, or the *fixed-interest* market in the UK and Commonwealth countries. Floating-rate notes (FRNs) do not have a fixed coupon at all but instead link their interest payments to an external reference, such as the three-month bank lending rate. Bank interest rates will fluctuate constantly during the life of the bond and so an FRN's cash flows are not known with certainty. Usually FRNs pay a fixed margin or *spread* over the specified reference rate; occasionally the spread is not fixed and such a bond is known as a *variable-rate note*. Because FRNs pay coupons based on the three-month or six-month bank rate they trade essentially as money market instruments.

I.B.2.2.2 Index-Linked Bonds

An index-linked bond has its coupon and redemption payment, or possibly just either one of these, linked to a specified index. When governments issue index-linked bonds the cash flows are

linked to a price index such as consumer or commodity prices. Corporations have issued index-linked bonds that are connected to inflation or a stock market index.

I.B.2.2.3 Zero-Coupon Bonds

Certain bonds do not make any coupon payments at all, and these are known as *zero-coupon bonds*. The only cash flow on a zero-coupon bond or *strip* is the redemption payment on maturity. If we assume that the maturity payment is, say, £100 or *par*, the issue price will be at a discount to par. Such bonds are also known therefore as *discounted* bonds. The difference between the price paid on issue and the redemption payment is the interest realised by the bondholder. This has certain advantages for investors, the main one being that there are no coupon payments to be invested during the bond's life. Both governments and corporations issue zero-coupon bonds. Conventional coupon-bearing bonds can be *stripped* into a series of individual cash flows, which would then trade as separate zero-coupon bonds. This is a common practice in government bond markets such as Treasuries or gilts, where the borrowing authority does not actually issue strips, and they have to be created via the stripping process.

I.B.2.2.4 Securitised Bonds

There is a large market in bonds whose interest and principal liability payments are backed by an underlying cash flow from another asset. By securitising the asset, a borrower can provide an element of cash-flow backing to investors. For instance, a mortgage bank can use the cash inflows it receives on its mortgage book as asset backing for an issue of bonds. Such an issue would be known as a mortgage-backed security (MBS). Because residential mortgages rarely run to their full term, but are usually paid off early by homeowners, the notes that are backed by mortgages are also prepaid ahead of their legal final maturity. This feature means that MBSs are not bullet bonds like vanilla securities, but are instead known as *amortising* bonds. Other asset classes that can be securitised include credit card balances, car loans, equipment lease receivables, nursing home receipts, museum or leisure park receipts and so on.

I.B.2.2.5 Bonds with Embedded Options

Some bonds include a provision in their offer particulars that gives the bondholder and/or the issuer an option to enforce early redemption of the bond. The most common type of option embedded in a bond is a *call feature*. A call provision grants the issuer the right to redeem all or part of the debt before the specified maturity date. An issuing company may wish to include such a feature as it allows it to replace an old bond issue with a lower coupon rate issue if interest rates in the market have declined. As a call feature allows the issuer to change the maturity date of a bond it is considered harmful to the bondholder's interests; therefore, the market price of the bond at any time will reflect this. A call option is included in all asset-backed securities based on

mortgages, for obvious reasons. A bond issue may also include a provision that allows the investor to change the maturity of the bond. This is known as a *put feature* and gives the bondholder the right to sell the bond back to the issuer at a predetermined price on specified dates. The advantage to the bondholder is that, if interest rates rise after the issue date, thus depressing the bond's value, the investor can realise par value by *putting* the bond back to the issuer. A *convertible* bond is an issue giving the bondholder the right to exchange the bond for a specified amount of shares (equity) in the issuing company. This feature allows the investor to take advantage of favourable movements in the price of the issuer's shares. The presence of embedded options in a bond makes valuation more complex than with plain vanilla bonds.

I.B.2.3 Pricing a Conventional Bond

The principles of pricing in the bond market are exactly the same as those in other financial markets, which state that the price of any financial instrument is equal to the net present value today of all the future cash flows from the instrument. A bond price is expressed as per 100 nominal of the bond, or 'per cent'. So, for example, if the 'all-in'² price of a US dollar-denominated bond is quoted as 98.00, this means that for every \$100 nominal of the bond a buyer would pay \$98. The interest rate or discount rate used as part of the present value (price) calculation is key to everything, as it reflects where the bond is trading in the market and how it is perceived by the market. All the determining factors that identify the bond – those discussed in this chapter and including the type of issuer, the maturity, the coupon and the currency – influence the interest rate at which a bond's cash flows are discounted, which will be roughly similar to the rate used for comparable bonds.

Since the price of a bond is equal to the present value of its cash flows, first we need to know the bond's cash flows before then determining the appropriate interest rate at which to discount the cash flows. We can then compute the price of the bond.

I.B.2.3.1 Bond Cash Flows

A vanilla bond's cash flows are the interest payments or coupons that are paid during the life of the bond, together with the final redemption payment. It is possible to determine the cash flows with certainty only for conventional bonds of a fixed maturity. So, for example, we do not know with certainty what the cash flows are for bonds that have embedded options and can be redeemed early. The coupon payments for conventional bonds are made annually, semi-annually or quarterly. Some bonds pay monthly interest.

² The 'all-in' price is the price actually paid for the bond in the market. See the discussion in Section I.B.2.3.6.

Therefore a conventional bond of fixed redemption date is made up of an annuity (its coupon payments) and the maturity payment. If the coupon is paid semi-annually, this means exactly half the coupon is paid as interest every six months. Both gilts and US Treasuries pay semi-annual coupons. For example, a 5% 2012 gilt has a semi-annual coupon = £100 × 0.025 = £2.50 and a redemption payment of £100. These are the only cash flows. The bond was issued on 23 June 1999 and is redeemed on 7 March 2012, and pays coupon on 7 March and 7 September each year. So in 2002 the bond is made up of 20 cash flows of £2.50 and one of £100. The time between coupon payments for any bond is counted as one period, so there are 20 periods between the first and last cash flows for the gilt in our example. The maturity payment is received 20 periods from today.

I.B.2.3.2 The Discount Rate

The interest rate that is used to discount a bond's cash flows (therefore called the *discount* rate) is the rate required by the bondholder. It is therefore known as the bond's *yield*. The yield on the bond will be determined by the market and is the price demanded by investors for buying it, which is why it is sometimes called the bond's *return*. The required yield for any bond will depend on a number of political and economic factors, including what yield is being earned by other bonds of the same class. Yield is always quoted as an annualised interest rate, so that for a semi-annually paying bond exactly half of the annual rate is used to discount the cash flows.

I.B.2.3.3 Conventional Bond Pricing

The *fair price* of a bond is the present value of all its cash flows. Therefore when pricing a bond we need to calculate the present value of all the coupon interest payments and the present value of the redemption payment, and sum these. The price of a conventional bond that pays annual coupons can therefore be given by:

$$\begin{aligned}
 P &= \frac{C}{1+r} + \frac{C}{(1+r)^2} + \frac{C}{(1+r)^3} + \dots + \frac{C}{(1+r)^N} + \frac{M}{(1+r)^M} \\
 &= \sum_{n=1}^N \frac{C}{(1+r)^n} + \frac{M}{(1+r)^N}
 \end{aligned}
 \tag{I.B.2.1}$$

where

- P is the price
- C is the annual coupon payment
- r is the discount rate (therefore, the required yield)
- N is the number of years to maturity (therefore, the number of interest periods in an annually paying bond; for a semi-annual bond the number of interest periods is $2N$)

M is the maturity payment or par value (usually 100% of currency).

The pricing process can be simplified by treating the coupon payments as an annuity stream and using the formula for valuation of an annuity. The present value of an annuity of \$1, where the payment is made at the end of each interest period is:

$$\frac{1 - 1/(1 + r)^N}{r}$$

The basic formula (I.B.2.1) can be modified in a number of ways. For example, the price of a bond that pays semi-annual coupons is given by an expression similar to (I.B.2.1), but modified to allow for the twice-yearly discounting: we set each payment at $C/2$ and use $2N$ as the power to which to raise the discount factor, as there are two interest payments every year for a bond that pays semi-annually. In the general expression it is therefore convenient to use the number of interest periods in the life of the bond, as opposed to the number of years to maturity, which we denote by as n .

Also, the basic formula calculates the fair price on a coupon payment date, so that there is no *accrued interest* incorporated into the price. That is, the standard price formula also assumes that the bond is traded for price settlement on a day that is precisely one interest period before the next coupon payment. More generally, the date used as the point for calculation is the *settlement date* for the bond, the date on which a bond will change hands after it is traded. For a new issue of bonds the settlement date is the day when the stock is delivered to investors and payment is received by the bond issuer. The settlement date for a bond traded in the *secondary market* is the day on which the buyer transfers payment to the seller of the bond and when the seller transfers the bond to the buyer. Different markets will have different settlement conventions: for example, UK gilts normally settle one business day after the trade date (the notation used in bond markets is ‘T + 1’), whereas Eurobonds settle on T + 3. The term *value date* is sometimes used in place of settlement date; however, the two terms are not strictly synonymous. A settlement date can fall only on a business day, so that a gilt traded on a Friday will settle on a Monday. However, a value date can sometimes fall on a non-business day, for example when accrued interest is being calculated.

The standard price formula is adjusted if dealing takes place between coupon dates. If we take the value date (almost always the settlement date, although unlike the settlement date the value date can fall on a non-working day) for any transaction, we then need to calculate the number of calendar days from this day to the next coupon date. We then use the following ratio i when adjusting the exponent for the discount factor:

$$i = \frac{\text{Days from value date to next coupon date}}{\text{Days in the interest period}}$$

The number of days in the interest period is the number of calendar days between the last coupon date and the next one, and it will depend on the day count basis used for that specific bond.³ The price formula is then modified as:

$$P = \frac{C}{(1+r)^i} + \frac{C}{(1+r)^{1+i}} + \frac{C}{(1+r)^{2+i}} + \dots + \frac{C}{(1+r)^{n-1+i}} + \frac{M}{(1+r)^{n-1+i}} \quad (\text{I.B.2.2})$$

where the variables C , M , n and r are as before. Note that (I.B.2.2) assumes r for an annually paying bond and is adjusted to $r/2$ for a semi-annually paying bond.

Example I.B.2.1

In these examples we illustrate the price calculation, using the annuity formula to simplify the calculation of the present value of the coupon payments.

(a) Calculate the fair pricing of a UK gilt, the 9% Treasury 2008, which pays semi-annual coupons, with the following terms: $C = 9\%$; $M = \text{£}100$; $N = 10$ years; $r = 4.98\%$

$$\begin{aligned} P &= 9.00 \left[\frac{1 - 1 / \left(1 + \frac{1}{2} \times 0.0498\right)^{20}}{0.0498 / 2} \right] + \frac{100}{\left(1 + \frac{1}{2} \times 0.0498\right)^{20}} \\ &= \text{£}70.2175 + \text{£}61.1463 = \text{£}131.3638 \end{aligned}$$

The fair price of the gilt is £131.3638, which is composed of the present value of the stream of coupon payments (£70.2175) and the present value of the return of the principal (£61.1463).

(b) What is the price of a 5% coupon sterling bond with precisely five years to maturity, with semi-annual coupon payments, if the yield required is 5.40%?

As the cash flows for this bond are 10 semi-annual coupons of £2.50 and a redemption payment of £100 in 10 six-month periods from now, the price of the bond can be obtained by solving the following expression, where we substitute $C = 2.5$, $n = 10$ and $r = 0.027$ into the price equation (the values for C and r reflect the adjustments necessary for a semi-annual paying bond):

$$P = 2.5 \left[\frac{1 - 1 / 1.027^{10}}{0.027} \right] + \frac{100}{1.027^{10}} = \text{£}21.65574 + \text{£}76.61178 = \text{£}98.26752$$

The price of the bond is £98.2675 per £100 nominal.

³ Day-count conventions are covered in Section I.B.1.2.

(c) *What is the price of a 5% coupon euro bond with five years to maturity paying annual coupons, again with a required yield of 5.40%?*

In this case there are five periods of interest, so we may set $C = 5$, $n = 5$, with $r = 0.05$.

$$\begin{aligned} P &= 5 \left[\frac{1 - 1/1.054^5}{0.054} \right] + \frac{100}{1.054^5} \\ &= \pounds 21.410121 + \pounds 76.877092 = \pounds 98.287213 \end{aligned}$$

Note how the bond paying annually has a slightly higher price for the same required annualised yield. This is because the sterling bond paying semi-annually has a higher effective yield than the euro bond, resulting in a lower price.

(d) *Consider our 5% sterling bond again, but this time the required yield has risen and is now 6%.*

This makes $C = 2.5$, $n = 10$ and $r = 0.03$.

$$\begin{aligned} P &= 2.5 \left[\frac{1 - 1/1.03^{10}}{0.03} \right] + \frac{100}{1.03^{10}} \\ &= \pounds 21.325507 + \pounds 74.409391 = \pounds 95.734898 \end{aligned}$$

As the required yield has risen, the discount rate used in the price calculation is now higher, and the result of the higher discount is a lower present value (price).

(e) *Calculate the price of our sterling bond, still with five years to maturity but offering a yield of 5.10%.*

$$\begin{aligned} P &= 2.5 \left[\frac{1 - 1/1.0255^{10}}{0.0255} \right] + \frac{100}{1.0255^{10}} \\ &= \pounds 21.823737 + \pounds 77.739788 = \pounds 99.563525 \end{aligned}$$

To satisfy the lower required yield of 5.10% the price of the bond has fallen to $\pounds 99.56$ per $\pounds 100$.

(f) *Calculate the price of the 5% sterling bond one year later, with precisely four years left to maturity and with the required yield still at the original 5.40%.*

This sets the terms in (b) unchanged, except now $n = 8$.

$$\begin{aligned} P &= 2.5 \left[\frac{1 - 1/1.027^8}{0.027} \right] + \frac{100}{1.027^8} \\ &= \pounds 17.773458 + \pounds 80.804668 = \pounds 98.578126 \end{aligned}$$

The price of the bond is $\pounds 98.58$. Compared with (b) this illustrates how, other things being equal, the price of a bond will approach par ($\pounds 100$ per cent) as it approaches maturity.

I.B.2.3.4 Pricing Undated Bonds

Perpetual or *irredeemable* bonds have no redemption date, so that interest on them is paid indefinitely. They are also known as *undated* bonds. An example of an undated bond is the 3.5% War Loan, a gilt formed out of issues in 1916 to help pay for the 1914–18 war effort. Most undated bonds date from a long time in the past and it is unusual to see them issued today. In structure the cash flow from an undated bond can be viewed as a continuous annuity or a perpetuity. The fair price of such a bond is given by letting $N \rightarrow \infty$ in (I.B.2.1) so that $P = C/r$, where the inputs C and r are as before.

I.B.2.3.5 Pricing Conventions

The convention in most bond markets is to quote prices as a percentage of par. The value of par is assumed to be 100 units of currency unless otherwise stated. A sterling bond quoted at an *offer* price of £98.45 means that £100 nominal of the bond will cost a buyer £98.45. A bond selling at below par is considered to be trading at a *discount*, while a price above par means the bond is trading at a *premium* to par. Do not confuse the term ‘trading at a discount’ with a discount instrument, however, which generally refers to a zero-coupon bond.

In most markets bond prices are quoted in decimals, in minimum increments of $1/100 = 0.01$. This is the case, for example, with Eurobonds, euro-denominated bonds and gilts. Certain markets, including the US Treasury market, for example, and certain Commonwealth markets such as South African and Indian government bonds, quote prices in *ticks*, where the minimum increment is $1/32$. One tick is therefore equal to 0.03125. A US Treasury might be priced at ‘98–05’ which means ‘98 and 5 ticks’. This is equal to $98 + 5/32 = 98.15625$.

Example I.B.2.2

What is the total consideration for £5 million nominal of a gilt, where the price is £114.50?

The price of the gilt is £114.50 per £100, so the consideration is:

$$1.145 \times 5,000,000 = \text{£}5,725,000$$

What consideration is payable for \$5 million nominal of a US Treasury, quoted at an all-in price of 99–16?

The US Treasury price is 99–16, which is equal to $99 + 16/32$, or 99.50 per \$100. The consideration is therefore: $0.9950 \times 5,000,000 = \$4,975,000$. If the price of a bond is below par the total consideration is below the nominal amount, whereas if it is priced above par the consideration will be above the nominal amount.

I.B.2.3.6 Clean and Dirty Bond Prices: Accrued Interest

Our discussion of bond pricing up to now has ignored coupon interest. All bonds (except zero-coupon bonds) accrue interest on a daily basis, and this is then paid out on the coupon date. The calculation of bond prices using present value analysis does not account for coupon interest or *accrued interest*. In all major bond markets the convention is to quote price as a *clean price*. This is the price of the bond as given by the net present value of its cash flows, but excluding coupon interest that has accrued on the bond since the last dividend payment. As all bonds accrue interest on a daily basis, even if a bond is held for only one day, interest will have been earned by the bondholder. However, we have referred already to a bond's *all-in price*, which is the price that is actually paid for the bond in the market. This is also known as the *dirty price* (or *gross price*), which is the clean price of a bond plus accrued interest. In other words, the accrued interest must be added to the quoted price to give the total consideration for the bond.

Accruing interest compensates the seller of the bond for giving up all of the next coupon payment even though they will have held the bond for part of the period since the last coupon payment. The clean price for a bond will move with changes in market interest rates; assuming that these are constant in a coupon period, the clean price will be constant for this period. However, the dirty price for the same bond will increase steadily from one interest payment date until the next. On the coupon date the clean and dirty prices are the same and the accrued interest is zero. Between the coupon payment date and the next *ex-dividend* date the bond is traded *cum dividend*, so that the buyer gets the next coupon payment. The seller is compensated for not receiving the next coupon payment by receiving accrued interest instead. This is positive and increases up to the next ex-dividend date, at which point the dirty price falls by the present value of the amount of the coupon payment. The dirty price at this point is below the clean price, reflecting the fact that accrued interest is now negative. This is because after the ex-dividend date the bond is traded 'ex-dividend'; the seller, not the buyer, receives the next coupon and the buyer has to be compensated for not receiving the next coupon by means of a lower price for holding the bond.

The net interest accrued since the last ex-dividend date is determined as follows:

$$AI = C \times \frac{N_{xt} - N_{xc}}{\text{Day base}} \quad (\text{I.B.2.3})$$

where

AI is the net accrued interest

C is the bond coupon

N_{xt} is the number of days between the ex-dividend date and the coupon payment date (7 business days for UK gilts)

N_{xc} is the number of days between the ex-dividend date and the date for the calculation

Day base is the day count base (usually 365 or 360).

Interest accrues on a bond from and including the last coupon date up to and excluding what is called the *value date*. The value date is almost always the *settlement* date for the bond, or the date when a bond is passed to the buyer and the seller receives payment. Interest does not accrue on bonds whose issuer has subsequently gone into default. Bonds that trade without accrued interest are said to be trading *flat* or *clean*. By definition therefore,

$$\text{Clean price of a bond} = \text{Dirty price} - AI.$$

For bonds that are trading ex-dividend, the accrued coupon is negative and would be subtracted from the clean price. The calculation is given by:

$$AI = -C \times \frac{\text{Days to next coupon}}{\text{Day base}} \quad (\text{I.B.2.4})$$

Certain classes of bonds, for example US Treasuries and Eurobonds, do not have an ex-dividend period and therefore trade cum dividend right up to the coupon date.

The accrued interest calculation for a bond is dependent on the day-count basis specified for the bond in question. We have already seen that when bonds are traded in the market the actual consideration that changes hands is made up of the clean price of the bond together with the accrued interest that has accumulated on the bond since the last coupon payment; these two components make up the dirty price of the bond. When calculating the accrued interest, the market will use the appropriate day-count convention for that bond.

Example I.B.2.3

(a) *The 7% Treasury 2006 gilt has coupon dates of 7 June and 7 December each year. £100 nominal of the bond is traded for value on 27 August 2002. What is the accrued interest on the value date?*

On the value date 81 days have passed since the last coupon date. Under the day-count convention system of actual/actual (which came into effect for gilts in November 1998) the accrued calculation uses the actual number of days between the two coupon dates, giving:

$$7 \times \frac{81}{183} \times 0.5 = 1.54918$$

(b) A purchaser buys £25,000 nominal of the 7% 2006 gilt for value on 27 August 2002, at a price of 102.4375. How much does he actually pay for the bond?

The clean price of the bond is 102.4375. The dirty price of the bond is $102.4375 + 1.55342 = 103.99092$. The total consideration is therefore $1.0399092 \times 25,000 = \text{£}25,997.73$.

(c) *A Norwegian government bond with a coupon of 8% is purchased for settlement on 30 July 2003 at a price of 99.50. Assume that this is seven days before the coupon date and therefore the bond trades ex-dividend. What is the all-in price?*

The accrued interest is $-8 \times 7/365 = -0.153424\%$. The all-in price is therefore $99.50 - 0.1534 = 99.3466$.

I.B.2.4 Market Yield

Just as there are many different types of bond and many different types of borrower, so there are different types of yield. The price of any bond will change in line with changes in required yield, so that straightaway we can see that it is not price changes that we are really interested in, but yields. It is a change in required yield that will drive a change in price. So in the bond market we are concerned with examining the determinants of market yields.

The main quoted yield in any market is the government bond yield. This is the yield on a domestic market's government bonds. The required yield on these bonds is mainly a function of the central bank's *base rate* or *minimum lending rate*, set by the government or central bank itself. Other factors will also impact the yield, including the relative size of the public-sector budget deficit and national debt as a percentage of the national product (usually measured as gross domestic product or gross national product); the economic policies that are adopted; and, of course, supply and demand for government bonds themselves. A change in any of these factors can and does affect government bond prices. While it is common to view government bonds as the safest credit for investors, this really only applies to the largest developed country markets. Bonds issued by certain countries within the Organisation for Economic Cooperation and Development (OECD), for example Greece, South Korea and Mexico, are not given the highest possible rating by credit analysts.

Bonds issued by non-sovereign borrowers will be priced off government bonds, which means that the yields required on them will be at some level above their respective government bond yields, if they are domestic currency bonds. Bond yields are often quoted as a *yield spread* over the equivalent government bond. This is known as the *credit spread* on a bond. A change in the required credit spread for any bond will affect the bond's price. Credit spreads will fluctuate for a variety of reasons, including when there is a change in the way the borrower is perceived in the market (such as a poor set of financial results by a corporation), which will affect the rate at

which the borrower can raise funds. Credit spreads can sometimes change because comparable bond yields change, as well as due to supply-and-demand factors and liquidity factors.

I.B.2.4.1 Yield Measurement

Bonds are generally traded on the basis of their prices but, because of the complicated patterns of cash flows that different bonds can have, they are generally compared in terms of their yields. This means that a market maker will usually quote a two-way price at which they will buy or sell a particular bond, but it is the yield at which the bond is trading that is important to the market maker's customer. This is because a bond's price does not actually tell us anything useful about what we are getting. Remember that in any market there will be a number of bonds with different issuers, coupons and terms to maturity. Even in a homogeneous market such as the gilt market, different gilts will trade according to their own specific characteristics. To compare bonds in the market, therefore, we need the yield on any bond, and it is yields that we compare, not prices. A fund manager, quoted a price at which they can buy a bond, will be instantly aware of what yield that price represents, and whether this yield represents fair value. So it is the yield represented by the price that is the important figure for bond traders.

The yield on any investment is the interest rate that will make the present value of the cash flows from the investment equal to the initial cost (price) of the investment. So mathematically the yield on any investment is the interest rate that satisfies our basic bond price equation introduced earlier as equation (I.B.2.1). But, as we have noted, there are other types of yield measure used in the market for different purposes. The most important of these are bond redemption yields, *spot* rates and *forward* rates. We will now discuss each type of yield measure and show how it is computed, and then discuss the relative usefulness of each measure.

I.B.2.4.2 Current Yield

The simplest measure of the yield on a bond is the *current yield*, also known as the *flat yield*, *interest yield* or *running yield*. The running yield is given by:

$$rc = \frac{C}{P} \times 100 \quad (\text{I.B.2.5})$$

where

- rc is the current yield
- C is the bond coupon
- P is the clean price of the bond.

Current yield ignores any capital gain or loss that might arise from holding and trading a bond and does not consider the time value of money. It essentially calculates the bond coupon income as a proportion of the price paid for the bond, and to be accurate would have to assume that the bond was more like an annuity than a fixed-term instrument.

The current yield is useful as a ‘rough-and-ready’ interest-rate calculation; it is often used to estimate the cost of or profit from a short-term holding of a bond. For example, if other short-term interest rates such as the one-week or three-month rates are higher than the current yield, holding the bond is said to involve a *running cost*. This is also known as *negative carry* or *negative funding*. The term is used by bond traders and market makers and *leveraged* investors. The *carry* on a bond is a useful measure for all market practitioners as it illustrates the cost of holding or funding a bond. The funding rate is the bondholder’s short-term cost of funds. A private investor could also apply this to a short-term holding of bonds.

Example I.B.2.4

(a) *A bond with a coupon of 6% is trading at a clean price of 97.89. What is the current yield of the bond?*

$$rc = \frac{6.00}{97.89} \times 100 = 6.129\%$$

(b) *What is the current yield of a bond with 7% coupon and a clean price of 103.49?*

$$rc = \frac{7.00}{103.49} \times 100 = 6.76\%$$

Note that the current yield of a bond will lie above the coupon rate if the price of the bond is below par, and vice versa if the price is above par.

I.B.2.4.3 Yield to Maturity

The *yield to maturity* (YTM) or *gross redemption yield* is the most frequently used measure of return from holding a bond. Yield to maturity takes into account the pattern of coupon payments, the bond’s term to maturity and the capital gain (or loss) arising over the remaining life of the bond. We saw from our bond price in equation (I.B.2.1) that these elements are all related and are important components determining a bond’s price. Indeed the formula for YTM is essentially that for calculating the price of a bond. For a bond paying annual coupons the YTM is calculated by solving the equation below, and we assume that the first coupon will be paid exactly one interest period from now (which for an annual coupon bond is exactly one year from now):

$$P_d = \frac{C}{(1+rm)^1} + \frac{C}{(1+rm)^2} + \frac{C}{(1+rm)^3} + \dots + \frac{C}{(1+rm)^n} + \frac{M}{(1+rm)^n} \quad (\text{I.B.2.6})$$

where

- P_d is the bond dirty price
- C is the coupon rate
- M is the par or redemption payment (100)
- rm is the annual YTM
- n is the number of interest periods.

For an annual coupon bond with maturity N years, equation (I.B.2.6) may be written as:

$$P_d = \sum_{n=1}^N \frac{C}{(1+rm)^n} + \frac{M}{(1+rm)^n}$$

For a semi-annual coupon bond we have:

$$P_d = \sum_{n=1}^N \frac{C/2}{(1+rm/2)^n} + \frac{M}{(1+rm/2)^n}$$

where N is now the number of interest periods in the life of the bond and therefore equal to the number of years to maturity multiplied by 2.

Note that the YTM expression has two variable parameters, the price P_d and yield rm . It cannot be rearranged to solve for yield rm explicitly and in fact the only way to solve for the yield is to use the process of numerical iteration.

Example I.B.2.5

For a five-year annual bond with coupon of 6% trading at a price of 97.89 we have:

$$\begin{aligned} \text{Running yield} &= 6.129\% \\ \text{Redemption yield} &= 6.50\% \end{aligned}$$

The YTM is equivalent to the *internal rate of return* (IRR) on the bond, the rate that equates the value of the discounted cash flows on the bond to its current dirty price. In effect both measures are identical; the assumption of uniform reinvestment rate allows us to calculate the IRR as equivalent to the rm . It is common for the IRR measure to be used by corporate financiers for project appraisal, while the YTM measure is used in bond markets.

Note that the YTM, that is, the redemption yield, is the *gross* redemption yield, the yield that results from payment of coupons without deduction of any withholding tax. The *net redemption yield* is obtained by multiplying the coupon rate C by $(1 - \text{marginal tax rate})$. The net yield is what will be received if the bond is traded in a market where bonds pay coupon net, which means net of a withholding tax. The net redemption yield is always lower than the gross redemption yield.

We have already alluded to the key assumption behind the YTM calculation, namely that the rate r_m remains stable for the entire period of the life of the bond. By assuming the same yield we can say that all coupons are reinvested at the same yield r_m . For the bond in Example I.B.2.5, this means that if all the cash flows are discounted at 6.5% they will have a total present value or NPV of 97.89. At the same time, if all the cash flows received during the life of the bond are reinvested at 6.5% until the maturity of the bond, the final redemption yield will be 6.5%. This is patently unrealistic since we can predict with virtual certainty that interest rates for instruments of similar maturity to the bond at each coupon date will not remain at 6.5% for five years.

In practice, however, investors require a rate of return that is equivalent to the price that they are paying for a bond, and the redemption yield is, to put it simply, as good a measurement as any. A more accurate measurement might be to calculate present values of future cash flows using the discount rate that is equal to the market's view on where interest rates will be at that point, known as the *forward* interest rate. However, forward rates are *implied* interest rates, and a YTM measurement calculated using forward rates can be as speculative as one calculated using the conventional formula. This is because the actual market interest rate at any time is invariably different from the rate implied earlier in the forward markets. We shall see later in this chapter how the *zero-coupon* interest rate is the true interest rate for any term to maturity. However, the YTM is, despite the limitations presented by its assumptions, the main measure of return used in the markets.

I.B.2.5 Relationship between Bond Yield and Bond Price

A fundamental property of bonds is that an upward change in the price results in a downward move in the yield, and vice versa. This is of course immediately apparent since the price is the present value of the cash flows; as the required yield for a bond decreases (increases), the present value and hence the price of the cash flow for the bond will increase (decrease). It also reflects the fact that for plain vanilla bonds the coupon is fixed. Therefore it is the price of the bond that will need to fluctuate to reflect changes in market yields.

When the coupon rate of a plain vanilla bond is equal to the market rate, the bond price will be par (100). If the required interest rate in the market moves above a bond's coupon rate at any

point in time, the price of the bond will adjust downwards in order for the bondholder to realise the additional return required. Similarly if the required yield moves below the coupon rate, the price will move up to equate the yield on the bond to the market rate. As a bond will redeem at par, the capital appreciation realised on maturity acts as compensation when the coupon rate is lower than the market yield.

Table I.B.2.1 shows the prices for a hypothetical 7% coupon bond, quoted for settlement on 10 August 2003 and maturing on 10 August 2008. The bond pays annual coupons on a 30/360 basis. The prices are calculated by inserting the required yield values into the standard formulae for a set of cash flows. We can calculate the present value of the annuity stream represented by the bond and the present value of the final maturity payment. Note that when the required yield is at the same level as the bond’s fixed coupon (in this case 7%) the price of the bond is 100%, or *par*.

Table I.B.2.1: Prices and yields for a 7% five-year bond

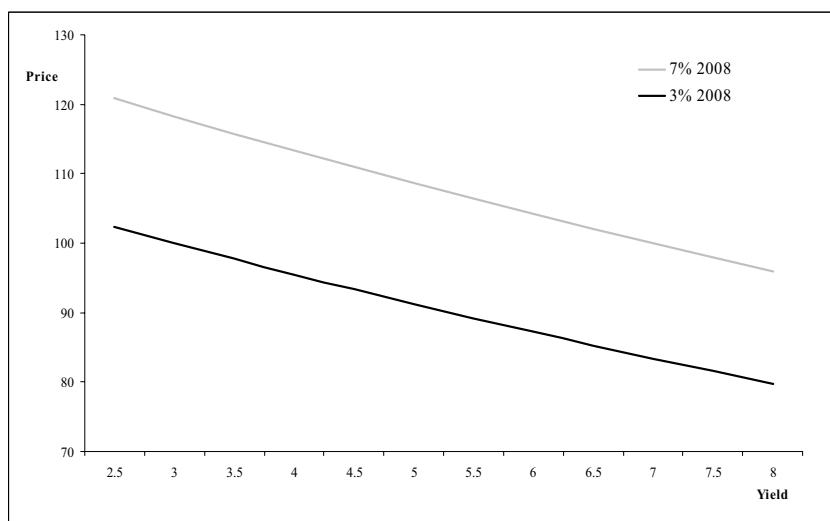
Yield (%)	Price
4.0	113.3555
4.5	110.9750
5.0	108.6590
5.5	106.4054
6.0	104.2124
6.5	102.0778
7.0	100.0000
7.5	97.9770
8.0	96.0073
8.5	94.0890
9.0	92.2207
9.5	90.4007
10.0	88.6276

The bond markets are also known as the ‘fixed-income’ or ‘fixed-interest’ markets. This reflects the fact that the coupon for conventional bonds is fixed, and in most cases the maturity date is also fixed. Therefore, when required yield levels in the market change, the price is the only factor that can change to reflect the new market yield levels. We saw in Table I.B.2.1 how the price of our hypothetical 7% five-year bond changed as the required yield changed. This is an important result. Let us consider this situation: if the required yield in the market for our 7% bond is fixed, investors will be happy to hold the bond. What if required yields subsequently rise above the 7% level? Bondholders will be unhappy, because they are now being paid 7% when elsewhere in the market higher yields are available. The market price of the bond will therefore change so that the yield on the bond changes, to compensate bondholders. If the required yield for the bond changes to 8%, we see from Table I.B.2.1 that the price will fall from par to 96.00. If this did not happen bondholders would sell the 7% issue and buy a bond that was yielding 8%; the market

price mechanism ensures that this does not happen. When required yield in the market is equal to a bond's coupon, the bond price will be par; the price will move respectively above or below par if required yields are below or above the coupon rate.

We illustrate the price/yield profile for two bonds of identical term to maturity but different coupons in Figure I.B.2.2. They are both from the same issuer, and hence of identical credit quality, with five years to maturity at the time the graph was plotted but with coupon levels of 3% and 7%. The higher coupon bond has the higher price for any given yield, but otherwise the price/yield profiles are very similar. However, as the bonds approach maturity, the convexity of the price/yield profile will increase, as we shall see in Section I.B.2.8 below. And the bond with the higher coupon, which will still have a higher price, will also have lower convexity.

Figure I.B.2.2: Price/yield profiles of two bonds with different coupons



If a bond is priced at below its par value it is said to be trading at a *discount*, while if it is trading above par value it is said to be trading at a *premium* to its par value. When investment banks issue bonds on behalf of borrowers, they will set a bond's coupon at the level that would make the bond price equal to par. This means that the bond coupon will be equal to the yield required by the market at the time of issue. The bond's price would then fluctuate as market yields changed, as we showed in Table I.B.2.1. Investors generally prefer to pay par or just under par when they buy a new issue of bonds, which is why an investment bank will set the coupon that equates the bond price to par or in a range between 99.00 and par. The reason behind this preference to pay no more than par is often purely cosmetic, since an issue price above par would simply indicate a coupon higher than the market rate. However, as many fund managers and investors buy a bond and hold it until maturity, one often finds this prejudice against paying over par for a new issue.

What will happen if market yields remain unchanged during the life of a bond from when it was issued? In this unlikely scenario the price of a conventional bond will remain unchanged at par. The price of any bond will ultimately equal its redemption value, which is par. Therefore a bond that is priced at a premium or a discount will gradually converge to par as it approaches maturity. This is sometimes referred to as the *pull-to-par phenomenon*.

As a bond approaches maturity there are fewer and fewer coupon payments, so that progressively more of the bond's price is made up of the present value of the final redemption payment. The present value of this payment will steadily increase as the maturity date is approached, since it is being discounted over a shorter period of time. The present value of the annuity cash flows of the bond steadily declines as we get fewer of them, and this is not offset by the increase in value of the maturity payment. Hence the price of the bond steadily declines. The opposite happens when the bond price starts off at a discount, where the increase in the value of the maturity payment outweighs the decrease in the price of the coupons, so that the bond price steadily converges to par.

I.B.2.6 Duration

Bonds pay a part of their total return during their lifetime, in the form of coupon interest, so that the term to maturity does not reflect the true period over which the bond's return is earned. Additionally, if we wish to gain an idea of the trading characteristics of a bond and compare this to other bonds of, say, similar maturity, the term to maturity is insufficient and so we need a more accurate measure. A plain vanilla coupon bond pays out a proportion of its return during the course of its life, in the form of coupon interest. If we were to analyse the properties of a bond, we should conclude quite quickly that its maturity gives us little indication of how much of its return is paid out during its life, nor any idea of the timing or size of its cash flows, and hence its sensitivity to moves in market interest rates. For example, if comparing two bonds with the same maturity date but different coupons, the higher coupon bond provides a larger proportion of its return in the form of coupon income than does the lower coupon bond. The higher coupon bond provides its return at a faster rate; its value is, theoretically, therefore less subject to subsequent fluctuations in interest rates.

We may wish to calculate an average of the time to receipt of a bond's cash flows, and use this measure as a more realistic indication of maturity. However, cash flows during the life of a bond are not all equal in value, so a more accurate measure would be to take the average time to receipt of a bond's cash flows, but weighted in the form of the cash flows' present value. This is, in effect, *duration*. We can measure the speed of payment of a bond, and hence its price risk relative to other bonds of the same maturity, by measuring the average maturity of the bond's cash flow

stream. Bond analysts use duration to measure this property (it is sometimes known as *Macaulay's duration*, after its inventor, who first introduced it in 1938).⁴ Duration is the weighted average time until the receipt of cash flows from a bond, where the weights are the present values of the cash flows, measured in years. At the time that he introduced the concept, Macaulay used the duration measure as an alternative for the length of time that a bond investment had remaining to maturity. A modified form of the Macaulay duration also provides a measure of the interest-rate risk from holding a bond.

I.B.2.6.1 Calculating Macaulay Duration and Modified Duration

Recall that the price/yield formula for a plain vanilla bond is (with the yield to maturity denoted by the symbol r in this section):

$$P = \frac{C}{1+r} + \frac{C}{(1+r)^2} + \frac{C}{(1+r)^3} + \dots + \frac{C}{(1+r)^n} + \frac{M}{(1+r)^n} \quad (\text{I.B.2.7})$$

If we take the first derivative of this we obtain

$$\frac{dP}{dr} = \frac{-1C}{(1+r)^2} + \frac{-2C}{(1+r)^3} + \dots + \frac{(-n)C}{(1+r)^{n+1}} + \frac{(-n)M}{(1+r)^{n+1}} \quad (\text{I.B.2.8})$$

giving the following equation to calculate the approximate *change in price for a small change in yield*:

$$\frac{dP}{dr} = -\frac{1}{(1+r)} \left[\frac{1C}{(1+r)} + \frac{2C}{(1+r)^2} + \dots + \frac{nC}{(1+r)^n} + \frac{nM}{(1+r)^n} \right] \quad (\text{I.B.2.9})$$

Readers may feel a sense of familiarity regarding the expression in brackets in equation (I.B.2.9), as this is the weighted average time to maturity of the cash flows from a bond, where the weights are the present values of each cash flow. If we divide both sides of (I.B.2.9) by P we obtain the expression for the approximate *percentage* price change:

$$\frac{dP}{dr} \frac{1}{P} = -\frac{1}{(1+r)} \left[\frac{1C}{(1+r)} + \frac{2C}{(1+r)^2} + \dots + \frac{nC}{(1+r)^n} + \frac{nM}{(1+r)^n} \right] \frac{1}{P} \quad (\text{I.B.2.10})$$

The bracketed expression in (I.B.2.10) divided by the current price of the bond P is the Macaulay duration, D :

$$D = \frac{1}{P} \left[\frac{1C}{(1+r)} + \frac{2C}{(1+r)^2} + \dots + \frac{nC}{(1+r)^n} + \frac{nM}{(1+r)^n} \right]$$

⁴ See Macaulay (1938). This remains a fascinating read and is available from Risk Classics publishing, under the title *Interest Rates, Bond Yields and Stock Prices in the United States since 1856*.

or

$$D = \frac{1}{P} \sum_{n=1}^N \frac{nC_n}{(1+r)^n} \quad (\text{I.B.2.11})$$

where C_n represents the bond cash flow at time n .

Example I.B.2.6

Calculate the Macaulay duration of a 10 year annual 8% coupon bond.

Period (n)	Cash flow	PV at current yield*	$n \times$ PV
1	8	7.43260	7.4326
2	8	6.90543	13.81086
3	8	6.41566	19.24698
4	8	5.96063	23.84252
5	8	5.53787	27.68935
6	8	5.14509	30.87054
7	8	4.78017	33.46119
8	8	4.44114	35.529096
9	8	4.12615	37.13535
10	108	51.75222	517.5222
Total		102.49696	746.540686

*Calculated as $C/(1+r)^n$.

Price = 102.497; Yield = 7.643%

Macaulay duration = $746.540686/102.497 = 7.283539998$ years.

The Macaulay duration value is measured in years. An interesting observation by Burghardt (1994) is that, ‘measured in years, Macaulay’s duration is of no particular use to anyone.’ This is essentially correct. However, as a risk measure and hedge calculation measure, duration transformed into *modified duration* was the primary measure of interest-rate risk used in the markets, and is still widely used despite the advent of the value-at-risk measure for market risk. We have

$$\frac{dP}{dr} \frac{1}{P} = -\frac{1}{1+r} D \quad (\text{I.B.2.12})$$

where D is the Macaulay duration. The modified duration, MD , is

$$MD = \frac{D}{1+r} \quad (\text{I.B.2.13})$$

If we are determining duration longhand, there is another arrangement we can use to shorten the procedure. Assuming an annual coupon bond priced on a date that leaves a complete number of years to maturity and with no interest accrued:

$$P = C \left[\frac{1 - 1/(1+r)^n}{r} \right] + \frac{M}{(1+r)^n}$$

This expression calculates the price of a bond as the present value of the stream of coupon payments and the present value of the redemption payment. If we take the first derivative of this and then divide by the current price of the bond P , the result is another expression for the modified duration:

$$MD = \frac{1}{P} \left(\frac{C}{r^2} \left[1 - \frac{1}{(1+r)^n} \right] + \frac{n(M - C/r)}{(1+r)^{n+1}} \right)$$

Example I.B.2.7

Calculate the modified duration of the bond in Example I.B.2.6.

The above expression becomes:

$$MD = \frac{1}{102.497} \left(\frac{8}{0.07634^2} \left[1 - \frac{1}{1.07634^{10}} \right] + \frac{10(100 - 8/0.07634)}{1.07634^{11}} \right) = 6.766947$$

Note that to obtain the Macaulay duration we multiply the modified duration by $(1 + r)$, in this case 1.07634, which gives us a value of 7.28354 years.

For an irredeemable bond, duration is given by $1/r_c$ where $r_c = (C/P_d)$ is the *running yield* (or *current yield*) of the bond. In general, duration increases with the maturity of the bond, with an upper bound given by the inverse of the running yield. For bonds trading at or above par, duration increases with maturity and approaches this limit from below. For bonds trading at a discount to par, duration increases to a maximum at around 20 years and then declines towards the inverse of the running yield.

I.B.2.6.2 Properties of the Macaulay Duration

A bond's duration is always less than its maturity. This is because some weight is given to the cash flows in the early years of the bond's life, which brings forward the average time at which cash flows are received. In the case of a zero-coupon bond, there is no present-value weighting of the cash flows, for the simple reason that there are no cash flows, and so the duration for a zero-coupon bond is equal to its term to maturity. Duration varies with coupon, yield and maturity.

The following three factors imply higher duration for a bond:

- the lower the coupon;
- the lower the yield;
- broadly, the longer the maturity.

Duration increases as coupon and yield decrease. As the coupon falls, more of the relative weight of the cash flows is transferred to the maturity date and this causes duration to rise. Because the coupon on index-linked bonds is generally much lower than on vanilla bonds, this means that the duration of index-linked bonds will be much higher than for vanilla bonds of the same maturity. As yield increases, the present values of all future cash flows fall, but the present values of the more distant cash flows fall relatively more than those of the nearer cash flows. This has the effect of increasing the relative weight given to nearer cash flows and hence of reducing duration.

The effect of the coupon frequency

Certain bonds such as Eurobonds pay coupon annually compared with, say, gilts which pay semi-annual coupons. If we imagine that every coupon is divided into two parts, with one part paid a half-period earlier than the other, this will represent a shift in weight to the left, as part of the coupon is paid earlier. Thus, increasing the coupon frequency shortens duration, and of course decreasing coupon frequency has the effect of lengthening duration.

Duration as maturity approaches

Using our definition of duration we can see that initially it will decline slowly, and then at a more rapid pace as a bond approaches maturity.

Duration of a portfolio

Portfolio duration is a weighted average of the duration of the individual bonds. The weights are the present values of the bonds divided by the full price of the entire portfolio, and the resulting duration calculation is often referred to as a ‘market-weighted’ duration. This approach is in effect the duration calculation for a single bond. Portfolio duration has the same application as duration for an individual bond, and can be used to structure an *immunised* portfolio.

I.B.2.6.3 Properties of the Modified Duration

Although it is common for newcomers to the market to think intuitively of duration much as Macaulay originally did, as a proxy measure for the time to maturity of a bond, such an interpretation misses the main point of duration, which is a measure of price volatility or interest-rate risk. From (I.B.2.12) and (I.B.2.13) we have:

$$\frac{dP}{dr} \frac{1}{P} = -MD \quad (\text{I.B.2.14})$$

Thus the modified duration is *the percentage change in the bond price for a 1% (absolute) increase in the yield*. (This explains why the term *volatility* is sometimes used to refer to modified duration, but this is becoming increasingly uncommon in order to avoid confusion with option markets' use of the same term, which there often refers to *implied volatility* and is something completely different.)

Example I.B.2.8

An 8% annual coupon bond is trading at par and has a (Macaulay) duration of 2.74 years. If yields rise from 8% to 8.50%, what will the new price of the bond be?

$$\Delta P = -D \times \frac{\Delta r}{1+r} \times P = -2.74 \times \frac{0.005}{1.080} \times 100 = -\pounds 1.2685$$

That is, the price of the bond will now be $\pounds 98.7315$. The modified duration of a bond with a duration of 2.74 years and yield of 8% is obviously $2.74/1.08 = 2.537$ years. This tells us that for a 1% move in the yield to maturity, the price of the bond will move (in the opposite direction) by 2.54%.

We can use modified duration to approximate bond prices for a given yield change:

Example I.B.2.9

For a bond with a modified duration of 3.99, priced at par, how much will the price fall if there is an increase in yield of one basis point?

Since 100 basis points equal 1%, $\Delta P = (-3.24/100) \times 0.01 \times 100 = \pounds 0.0399$, or 3.99 pence.

I.B.2.7 Hedging Bond Positions

In Example I.B.2.9, 3.99 pence was the *basis-point value* (BPV) of the bond, that is, the change in the bond price given a one-basis-point change in the bond's yield. The basis-point value of a bond is:

$$BPV = \frac{MD}{100} \times \frac{P}{100} \tag{I.B.2.15}$$

Basis-point values are used in hedging bond positions. To hedge a bond position requires an opposite position to be taken in the hedging instrument. So if we are long a 10-year bond, we may wish to sell short a similar 10-year bond as a hedge against it. Similarly, a short position in a bond will be hedged through the purchase of an equivalent amount of the hedging instrument. In fact there is a variety of hedging instruments available, both on and off balance sheet.

Once the hedge is put on, any loss in the primary position should in theory be offset by a gain in the hedge position, and vice versa. The objective of a hedge is to ensure that the price change in the primary instrument is equal to the price change in the hedging instrument. If we are hedging a position with another bond, we use the BPVs of each bond to calculate the amount of the hedging instrument required. This is important because each bond will have different BPVs, so that to hedge a long position in, say, £1 million nominal of a 30-year bond does not mean we simply sell £1 million of another 30-year bond. This is because the BPVs of the two bonds will almost certainly be different. Also there may not be another 30-year bond available.

What if we have to hedge with a 10-year bond? How much nominal of this bond would be required? To calculate the nominal hedge position we need to know the ratio

$$\frac{BPV_p}{BPV_b}$$

where

BPV_p is the basis-point value of the primary bond (the position to be hedged)

BPV_b is the basis-point value of the hedging instrument.

The *hedge ratio* is used to calculate the size of the hedge position and is given by

$$\frac{BPV_p}{BPV_b} \times \frac{\text{change in yield for primary bond position}}{\text{change in yield for hedge instrument}} \quad (\text{I.B.2.16})$$

The second ratio in (I.B.2.16) is known as the *yield beta*.

Example I.B.2.10

A trader holds a long position of £1 million of the 8% 2019 bond. The modified duration of the bond is 11.14692 and its price is 129.87596. The basis-point value of this bond is therefore 0.14477. The trader decides to protect against a rise in interest rates, to hedge the position using the 0% 2009 bond, which has a BPV of 0.05549. If we assume that the yield beta is 1, what nominal value of the zero-coupon bond must be sold in order to hedge the position?

The hedge ratio is:

$$\frac{0.14477}{0.05549} \times 1 = 2.60894$$

Therefore to hedge £1 million of the 20-year bond the trader shorts £2,608,940 of the zero-coupon bond. If we use the respective BPVs to see the net effect of a one-basis-point rise in yield, the loss on the long position is approximately equal to the gain in the hedge position.

For a one-basis-point change in yield, the change in price given by the dollar duration figure, while not completely accurate, is a reasonable estimation of the actual change in price. For a large move, however, say 200 basis points, the approximation is significantly in error and analysts would not use it. The dollar duration value underestimates the change in price resulting from a large fall in yields but overestimates the price change for a large rise in yields. This is a reflection of the price/yield relationship for this bond. Some bonds will have a more pronounced convex relationship between price and yield, and the modified duration calculation will underestimate the price change resulting from either a fall or a rise in yields.

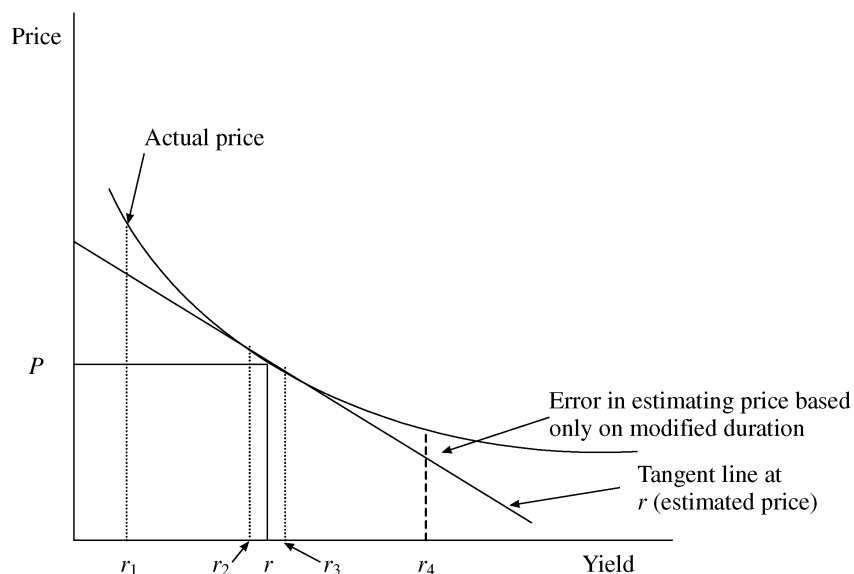
I.B.2.8 Convexity

Duration can be regarded as a first-order measure of interest-rate risk: it measures the *slope* of the present value/yield profile. It is, however, only an approximation for the actual change in bond price given a small change in yield to maturity. Similarly for modified duration and basis-point value, both of which describe the price sensitivity of a bond to small changes in yield. However, as Figure I.B.2.3 illustrates, the approximation is an underestimate of the actual price at the new yield. This is the weakness of the duration measure.

Convexity is a second-order measure of interest-rate risk; it measures the *curvature* of the present value/yield profile. Convexity can be regarded as an indication of the error we make when using duration and modified duration, as it measures the degree to which the curvature of a bond's price/yield relationship diverges from the straight-line estimation. The convexity of a bond is positively related to the dispersion of its cash flows. Thus, other things being equal, if one bond's cash flows are more spread out in time than another's, it will have a higher *dispersion* and hence a higher convexity. Convexity is also positively related to duration.

Figure I.B.2.3: Approximation of the bond price change using modified duration.

Reproduced with permission from Frank J Fabozzi (2002).



Convexity and its relationship to duration can best be understood by examining derivatives of the bond price formula with respect to yield. The difference of the bond price per unit change in r ($\Delta P/\Delta r$) is the slope of the price/yield function. The second difference ($\Delta^2 P/\Delta r^2$) tells how the slope is changing with yield. Whereas modified duration gives a rough approximation to the question of how price changes with yield, the first two derivatives can be combined in a Taylor series expansion to give a superior approximation:⁵

$$\Delta P = \frac{\Delta P}{\Delta r} \Delta r + \frac{1}{2} \frac{\Delta^2 P}{\Delta r^2} (\Delta r^2)$$

The expression above can be divided throughout by P to get the percentage price change that would result from a given change in the redemption yield r :

$$\frac{\Delta P}{P} = \frac{1}{P} \frac{\Delta P}{\Delta r} \Delta r + \frac{1}{2P} \frac{\Delta^2 P}{\Delta r^2} (\Delta r^2) \tag{I.B.2.17}$$

Note that the modified duration (MD) is contained within the first term on the right-hand side of equation (I.B.2.17). Market participants use the term convexity (CV) to refer to the second

⁵ Taylor series expansions were proposed by the English mathematician Brook Taylor in the eighteenth century as a method for approximating changes in complex functions. In this example we have used only the first two terms of the Taylor expansion using the first and second derivatives. This gives a ‘quadratic’ approximation, so named because it is a function of the change in yield squared. We could produce ever more sophisticated estimates of the change in price by introducing higher-order derivatives to give a cubic approximation, a quartic approximation etc.

derivative divided by the bond price. Note that CV is contained within the second term of equation (I.B.2.17), so we can rewrite (I.B.2.17) as follows:

$$\frac{\Delta P}{P} = -MD(\Delta r) + \frac{CV}{2}(\Delta r^2) \quad (\text{I.B.2.18})$$

Convexity is the rate at which price variation to yield changes with respect to yield. That is, it describes how a bond's modified duration changes with respect to changes in yield. Convexity can be approximated by⁶

$$CV = 10^8 \left(\frac{\Delta P'}{P} + \frac{\Delta P''}{P} \right) \quad (\text{I.B.2.19})$$

where $\Delta P'$ is the change in bond price if yield increases by one basis point, and $\Delta P''$ is the change in bond price if yield decreases by one basis point. Note that the value for convexity will be positive for a standard bond, that is, the approximate price change due to convexity is positive for both yield increases and decreases.

The convexity measure for a zero-coupon bond is given by:

$$CV = \frac{N(N+1)}{(1+r)^2} \quad (\text{I.B.2.20})$$

We discussed how the slope of the price/yield profile will change more rapidly for a bond of higher convexity, and that such a bond will outperform a bond of lower convexity whatever happens to market interest rates. High convexity is therefore a desirable property for bonds to have. In principle a more convex bond should fall in price less than a less convex one when yields rise, and rise in price more when yields fall. That is, convexity can be equated with the potential to outperform. Thus, other things being equal, the higher the convexity of a bond the more desirable it should, in principle, be to investors. In some cases investors may be prepared to accept a bond with a lower yield in order to gain convexity.

⁶ A more accurate method for calculating convexity is to divide the second derivative of the price equation by the price of the bond. For a standard bond with semi-annual coupons the second derivative of the price equation is:

$$\frac{\Delta^2 P}{\Delta r^2} = \sum_{n=1}^N \frac{n(n+1)C}{(1+r)^{n+2}} + \frac{N(N+1)M}{(1+r)^{N+2}} \quad (\text{I.B.2.21})$$

The unit of measurement for convexity using this method is the length of an interest period. For annual coupon bonds this is equal to the number of years; for bonds with semi-annual coupons it will be half-years. To express convexity in years we divide by the number of coupons per year squared. For example, in the case of a bond paying semi-annual coupons, convexity calculated using (I.B.2.21) would be expressed in half-years. To express in years, we would divide by 4.

What level of premium will be attached to a bond’s higher convexity? This is a function of the current yield levels in the market as well as market volatility. Both modified duration and convexity are functions of yield level, and so the effect of both is magnified at lower yield levels. As well as the relative level, investors will value convexity higher if the current market conditions are volatile. This is because the cash effect of convexity is noticeable only for large moves in yield. If an investor expects market yields to move only by relatively small amounts, they will attach a lower value to convexity; and vice versa for large movements in yield. Therefore, the yield premium attached to a bond with higher convexity will vary according to market expectations of the future size of interest-rate changes.

Example I.B.2.11

A four-year bond pays semi-annual coupons at a rate of 6%. If the yield is 6% p.a., its modified duration is 3.51 and its price is 100.00, what is the convexity of the bond expressed in half-years? How can we interpret this information?

The convexity can be approximated using equation (I.B.2.19). If the yield increases to 6.01% the price of the bond will decrease to 99.96490894. If the yield decreases to 5.99% the price of the bond will increase to 100.03510587. Therefore

$$CV = 10^8 \left(\frac{99.96490894 - 100.00}{100.00} + \frac{100.03510587 - 100.00}{100.00} \right) = 14.81$$

Note that the convexity may also be measured using the method explained in footnote 6 as follows:

Period, n	Cash Flow \$	$\frac{1}{(1.03)^{n+2}}$	$n(n+1)$ Cash Flow	$\frac{n(n+1)C}{(1.03)^{n+2}}$
1	3.00	0.915142	6	5.49
2	3.00	0.888487	18	15.99
3	3.00	0.862609	36	31.05
4	3.00	0.837484	60	50.25
5	3.00	0.813092	90	73.18
6	3.00	0.789409	126	99.47
7	3.00	0.766417	168	128.76
8	103.00	0.744094	7416	5518.20
Sum				5922.39

$$\text{Second derivative} = 5922.39; \text{Convexity (half-years)} = 5922.39/100.00 = 59.22$$

$$\text{Convexity (years)} = 59.22/4 = 14.81$$

Convexity tells us the size of the error that will arise when using modified duration to estimate changes in the price of a bond. For example, suppose that we wish to estimate the impact of an increase in yield by 1% (i.e. from 6.00% to 7.00%). Using modified duration alone, the estimated percentage change in price is equal to: $\Delta P/P = -3.51 \times 0.01 = -3.51\%$. This estimate of percentage price change using modified duration only overstates the adverse impact of the yield increase. A truer estimate can be found by using equation (I.B.2.18) and incorporating convexity as follows:

$$\Delta P/P = (-3.51 \times 0.01) + (\frac{1}{2} \times 14.81 \times 0.01^2) = -3.44\%$$

The usefulness of this new approximation can be assessed by comparing it to the *actual* change in price. What is the exact impact on the bond price of an increase in its required yield to 7%? We use conventional bond pricing techniques as follows:

$$P = 3.0 \left[\frac{1 - 1/1.035^8}{0.035} \right] + \frac{100}{1.035^8} = 96.563022$$

If the required yield increases to 7%, the bond price will fall to 96.563022, a fall of 3.44%. In this case the approximation provided by duration and convexity (-3.44%) gives an accurate estimate of the actual price movement.

We conclude this section with a brief discussion of the properties of convexity. That is, how does convexity vary according to the features of the bond?

- *Yield.* Yield and convexity are negatively correlated. As explained earlier, modified duration will decrease as yield increases. Positive convexity means that the rate of change in modified duration becomes smaller as yield increases.
- *Maturity.* From equation (I.B.2.21) we can see that the convexity measure increases with the square of maturity. Intuitively, this makes sense since longer-term bonds have more dispersed cash flows, thus creating greater sensitivity to changes in yield.
- *Coupon.* Applying the same logic, reducing the coupon rate also creates greater dispersion of cashflows, leading to higher convexity.
- *Index-linked bonds.* Index-linked bonds generally have much lower coupons than do vanilla bonds. Consistent with the point above, the lower coupon implies higher convexity.
- *Callable bonds.* A fall in interest rates increases the probability that the borrower will call the bonds in order to refinance. The potential for a price increase in the bond is therefore

truncated by this right to call. As a result, callable bonds and other bonds with prepayment features can have negative convexity, unlike their vanilla counterparts.

I.B.2.9 A Summary of Risks Associated with Bonds

The most important risk associated with bond investment and trading is *interest-rate risk*, a type of market risk. A change in the required yield for a particular bond, which is most often caused by a change in the market interest rate (the required yield on all comparable bonds), induces a change in the bond price (in the opposite direction, as shown in Section I.B.2.5). Thus interest-rate risk is typically a function of economic conditions, including inflation rates. The sensitivity of a bond to interest-rate risk is most often measured by its duration (see Section I.B.2.6) and its convexity (see Section I.B.2.8).

Figure I.B.2.4: Bloomberg YA screens for the two bonds in Figure I.B.2.2

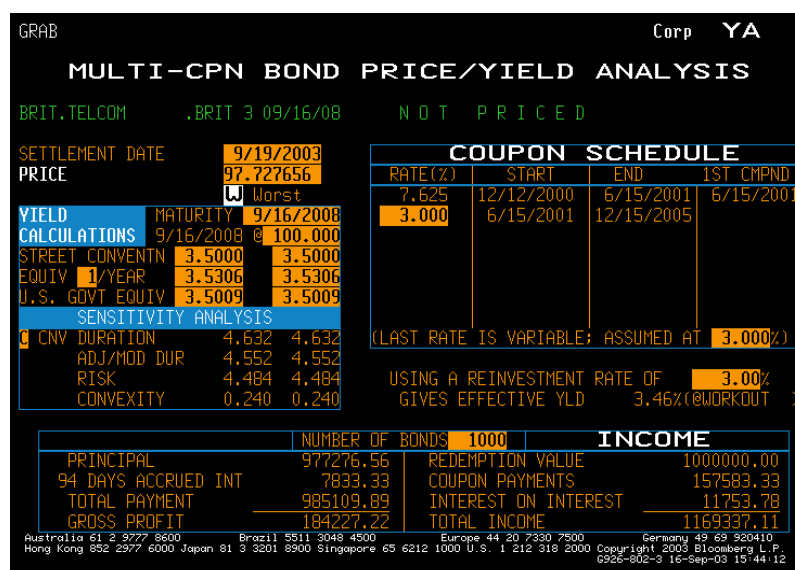
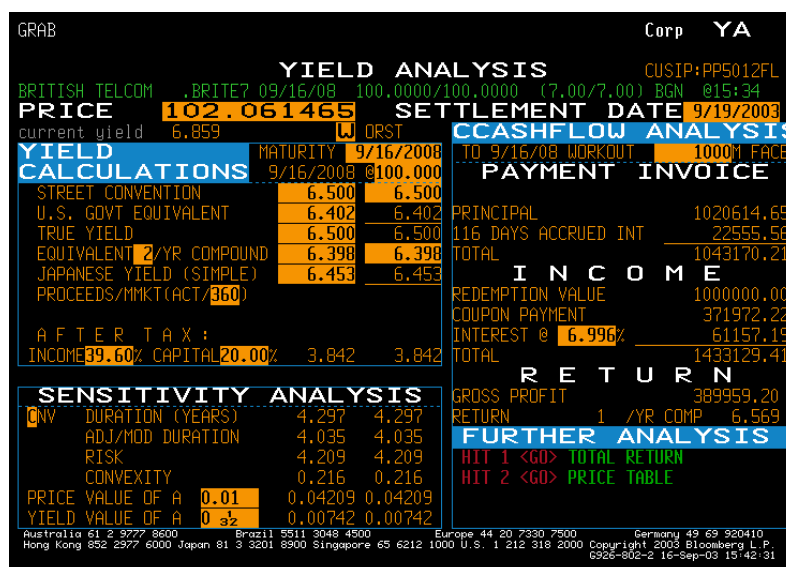


Figure I.B.2.4 shows Bloomberg screen YA (the yield analysis page) for each of the bonds whose price/yield relationship was shown in Figure I.B.2.2. The relationship profile in Figure I.B.2.2 was essentially identical for both bonds, except the bond with the higher coupon has a higher price for any given yield.

On Bloomberg screen YA, the price and/or yield may be entered by the user in the relevant fields. We see from the top screen in Figure I.B.2.4 that a yield of 6.5% has been entered for the 7% coupon bond, giving a price of 102.06. Similarly, for the 3% coupon bond in the lower screen a yield of 3.5% gives a price of 97.73. Both screens also give the Macaulay duration, the modified duration, and the ‘risk’ – each of these being a measure of the price/yield sensitivity. The ‘risk’ is simply the BPV, that is, the change in price for a one-basis-point change in yield. The reader will find it useful to obtain the figures given for duration and ‘risk’ by applying formulae (I.B.2.11), (I.B.2.13) and (I.B.2.15) to these bonds.

We have seen that the duration and BVP are linear approximations of a nonlinear relationship, the price/yield relationship, and the measure of their inaccuracy is given by the convexity. Note how small the convexity is for both these bonds – as expected from the nearly linear price/yield relationships shown in Figure I.B.2.2. Also note that the bond with lower coupon and yield has a higher convexity, consistent with the remarks made at the end of Section I.B.2.8.

The price of a bond will move for a variety of other reasons. Some of the other risk factors associated with bonds are noted here:

- *Credit risk.* A change in the yield required by the market may result from a perceived change in the credit quality of the bond issuer. However, credit considerations do not normally affect developed-country government bonds.
- *Liquidity risk.* Bond prices also move for liquidity reasons and normal supply-and-demand reasons. For example, if there is a large amount of a particular bond on issue it is easier to trade the bond; also if there is high demand due to a large customer base for the bond. Liquidity is a general term used here to mean the ease with which a market participant can trade in or out of a position (see Chapter I.C.1). If there is always a ready buyer or seller for a particular bond, it will be easier to trade in the market.
- *Reinvestment risk.* As noted in Section I.B.2.4, the yield-to-maturity calculation assumes that it will be possible to reinvest coupons at the same yield. Changes in this reinvestment rate will affect the realised yield.
- *Currency risk.* If an investor purchases a bond denominated in a currency other than his/her home currency, the investment will be exposed to exchange-rate movements.

- *Call risk.* If call provisions exist, the investor faces the risk that the bond will be called. Uncertainty exists with regard to the exact timing of the call and the reinvestment rate that will apply to the proceeds if the bond is called.
- *Volatility risk.* This type of risk is relevant for bonds with embedded options whose value is a function of interest-rate volatility.
- *Political/legal risk.* Governments may change tax laws or market regulations which may affect the value of a bond.

References

- Bierwag, G O (1978) 'Measures of duration', *Economic Inquiry*, 16, October, pp. 497–507.
- Burghardt, Galen (1994) *The Treasury Bond Basis*, p. 90. (Irwin, New York)
- Choudhry, M (2001) *The Bond and Money Markets: Strategy, Trading, Analysis* (Oxford: Butterworth-Heinemann), Chapters 2–10.
- Fabozzi, F (2002) *Bond Markets, Analysis and Strategies* 4th Edition (Wiley, New York), Chapter 2.
- Garbade, K (1996) *Fixed Income Analytics* (Cambridge, MA: MIT Press), Chapters 3, 4 and 12.
- Macaulay, F (1938) *Some Theoretical Problems Suggested by the Movements of Interest Rates, Bond Yields and Stock Prices in the United States since 1865* (New York: National Bureau of Economic Research).

I.B.3 Futures and Forwards

Keith Cuthbertson and Dirk Nitzsche¹

Forward and futures contracts are widely used for hedging and speculation. As we have seen in Chapter I.A.7, the cost-of-carry relationship ensures that the futures price and the spot price are closely linked. This high positive correlation between the two enables futures to be used effectively in speculation and hedging.

This chapter is organised as follows. First we provide a broad overview of the similarities and differences between forward and futures contracts. Then we discuss the way specific futures contracts on stock indices, currencies, commodities and fixed-income securities (i.e. T-bills, Eurodollar deposits/loans and T-bonds) are used in hedging the underlying spot-cash assets. The formulae used in calculating the optimal number of futures contracts to hedge an underlying spot (or cash) market position are very similar, although there are small differences due to the different contract specifications. Throughout we emphasise the intuitive ideas behind the various hedging strategies and provide a number of real-world practical examples.

Two other chapters of *The PRM Handbook* also deal with futures and forward contracts: the principles that are applied to the valuation of forward contracts are explained in Chapter I.A.7, and for discussion of the operation and characteristics of futures markets please refer to Chapter I.C.6.

I.B.3.1 Introduction

Analytically, ‘forwards’ and ‘futures’ are very similar, even though the contracts are traded differently in some respects. A holder of a long (short) forward contract has an agreement to buy (sell) an asset at a certain time in the future for a certain price, which is fixed today. The *buyer* (or *seller*) in a *forward* or *futures* contract:

- acquires a legal obligation to buy (or sell) an asset (the *underlying*)
- at some specific future date (*expiry date*)
- in an amount (the *contract size*);
- and at a price (the *forward* or *futures price*) which is fixed today.

A forward contract is an over-the-counter (OTC) instrument, and trades take place directly (usually over the phone) for a specific amount and specific delivery date as negotiated between

¹Keith Cuthbertson is Professor of Finance and Dirk Nitzsche is Senior Lecturer at the Cass Business School, City University, London.

the two parties. In contrast, *futures contracts* are standardised (in terms of contract size and delivery dates). Trades take place on an organised exchange and the contracts are revalued (marked to market) daily. When you buy or sell a futures contract on, say, cocoa it is the *legal right* to terms of the contract that is being purchased or sold, not the cocoa itself (which is actually bought and sold in the spot market for cocoa). As we have seen in Chapter I.A.7, there is a close link between the futures price and the spot price (for cocoa) but they are not the same thing!

Futures contracts are often traded between market makers in a ‘pit’ on the floor of an exchange, of which the largest are the Chicago Board of Trade (CBOT), the Chicago Mercantile Exchange (CME) and the Philadelphia Stock Exchange. However, in recent years there has been a move away from trading by ‘open outcry’ in a ‘pit’ towards electronic trading between market makers (and even more recently over the World Wide Web). For example, the London International Financial Futures Exchange (LIFFE) is now an electronic trading system, as are the European derivatives markets such as the French MATIF and EUREX (formerly DTB and SOFFEX, the German and Swiss derivatives exchanges).

A key feature of futures is that the contract calls for *deferred delivery* of the underlying asset (e.g. live hogs, cocoa), whereas spot assets are for *immediate delivery* (although in practice there is usually a delay of up to a few days). To distinguish between purchases and sales of futures contracts and the underlying (spot) asset, the latter are often referred to as transactions in the *cash* or *spot market*.

Today there are a large number of exchanges that deal in futures contracts, and most can be categorised as agricultural futures contracts (where the underlying ‘asset’ is, for example, pork bellies, live hogs or wheat), metallurgical futures (e.g. silver) or financial futures contracts (where the underlying asset could be a portfolio of stocks represented by the S&P 500, currencies, T-bills, T-bonds, Eurodollar deposits, etc.). Futures contracts in agricultural commodities have been traded (e.g. on CBOT) for over 100 years. In 1972 the CME began to trade currency futures, while interest-rate futures and stock index futures (‘pinstripe pork bellies’) were introduced in 1975 and 1982 respectively. The CBOT introduced a clearing house in 1925, where each party to the contract had to place ‘deposits’ into a margin account. This provides insurance if one of the parties defaults on the contract. The growth in the volume of futures trading since 1972 has been astounding: from 20 million contracts per year in 1972 to over 200 million contracts in the 1990s on the US markets alone. (For further details see Chapters I.C.6–8.)

Forwards and futures contracts differ in some practical details (see Table I.B.3.1). Forward contracts (usually) involve no ‘upfront’ payment and ‘cash’ changes hands only at the expiry of the contract. A forward contract is negotiated between two parties and is not marketable. In

contrast, a futures contract is traded in the market and involves a ‘down payment’ known as the *initial margin*. However, the initial margin is primarily a deposit to ensure both parties to the contract do not default. It is not a payment for the futures contract itself. The margin usually earns a competitive interest rate, so it is not a ‘cost’. As the futures price changes, payments are made into (or out of) the margin account. Hence a futures contract is a forward contract that is ‘marked to market’, daily.

Table I.B.3.1: Forward and futures contracts

Forwards	Futures
<ul style="list-style-type: none"> • Private (non-marketable) contract between two parties • (Large) trades are not communicated to other market participants • Delivery or cash settlement at expiry • Usually one delivery date • No cash paid until expiry • Negotiable choice of delivery dates, size of contract 	<ul style="list-style-type: none"> • Traded on an exchange • Trades are immediately known by other market participants • Contract is usually closed out prior to maturity • Range of delivery dates • Cash payments into (out of) margin account, daily • Standardised contract

Because the futures contract is marketable, the contracts have to be standardised: for example, by having a set of fixed expiry (delivery) dates and a fixed contract size (e.g. \$100,000 for the US T-bond futures on the International Money Market (IMM) division of the CME). In contrast, a *forward contract* can be ‘tailor-made’ between the two parties to the contract, in terms of ‘size’ and delivery date. Finally, forward contracts almost invariably involve the delivery of the underlying asset (e.g. currency), whereas futures contracts can be (and usually are) closed out by selling the contract prior to maturity. When they are, the clearing house sends out a cash payment that reflects the amount remaining in your margin account after all the daily adjustments have been made for gains and losses over the time you have held the contract. Because the price of a futures contract is derived from the price of the underlying spot asset, the changes in the futures price usually move (nearly) one-for-one with changes in the price of the underlying asset.

Speculation with futures is straightforward. Suppose you purchased a three-month futures contract at a price $F_0 = \$100$ and one month later you closed out the contract by selling it at the market price of $F_1 = \$110$. Then the clearing house ‘effectively’ sends you a cheque for the difference (\$10) and obtains this sum from the ‘short’, i.e. from the person who initially sold you the contract. (The institutional details differ from this, as we shall see in the Chapter I.C.6, but the principle is correct.) It also follows that you would earn a \$10 profit if you initially sold a contract at $F_0 = \$100$ and later closed out the contract by buying it back at $F_1 = \$90$ (‘sell high, buy low’). The possible types of futures contract that can be traded are almost limitless, but only

those that are useful for hedging and speculation will survive. The exchange will remove any futures contract where trading volume is low.

I.B.3.2 Stock Index Futures

Stock index futures are contracts traded on an underlying stock market index such as the S&P 500, FTSE 100 or Nikkei 225. Such futures contracts are widely used in hedging, speculation and index arbitrage. In a well-diversified portfolio of stocks all idiosyncratic risk of individual stocks has been eliminated and only market ('systematic' or 'non-diversifiable') risk remains (see Chapter I.A.3). Stock index futures can be used to eliminate the market risk of a portfolio of stocks. For example, a pension fund manager might fear a fall in equity prices over the next three months, which is the end of her evaluation period. She can use stock index futures to eliminate this market risk and hence she will effectively earn the risk-free rate of interest over the three-month period. She could also earn the risk-free-rate by selling all her stocks and holding T-bills. However, this would involve high transaction costs as she sold and then bought back the stocks three months later. Using index futures is far cheaper.

It is also the case that, if the fund manager believed the market would become increasingly volatile over the next three months (i.e. could go up or down a lot) and she wished to remove this uncertainty, she could again use index futures to eliminate the market risk. At the end of the three-month period, if she believes the stock market has returned to 'normal', she would close out her futures position and her stock portfolio would again be subject to market risk (for which she would receive higher expected returns than the risk-free rate). Alternatively, suppose you plan to invest cash that you will receive in three months' time in a diversified portfolio of stocks but you are worried that the market will rise over the next three months. Then you can lock in the known futures price today, by buying stock index futures and hedge against any future rise in the cash price of the stocks.

Speculation on a bear or bull market is easier to achieve with index futures rather than incurring the cost of buying or selling a large number of stocks that make up your prospective portfolio. Finally, if the quoted futures price is not equal to the synthetic futures price (see Chapter I.A.7) then you can make a (near) riskless arbitrage profit by simultaneously trading in the spot and futures markets on the index.

I.B.3.2.1 Contract Specifications

Stock index futures contracts are written on aggregate stock market indices and are settled in cash. In the USA index futures contracts traded on the CME include those on the S&P 500, the E-Mini-S&P 500, the S&P 500/BARRA GROWTH index and the Nikkei 225, as well as several

other indices. Trading volume is highest in the S&P 500 futures contract and, taking all of these contracts together, the dollar volume of the underlying stocks traded in these index futures contracts exceeds the dollar volume on the stock market itself.

On LIFFE futures contracts are available on various UK stock indices, including the FTSE 100 and FTSE 250, on European indices such as the FTSE Eurotop 100 and FTSE Eurotop 300 (including and excluding the UK), and several others. The reason for so many different futures contracts is so that investment managers can effectively hedge their equity portfolios, using futures with an underlying index that most closely matches the composition of the portfolio itself.

All stock index futures are settled in cash (with settlement procedures differing between the different contracts) but nearly all contracts are closed out prior to expiry. Cash settlement is based on the *value of an index point* $\$z$ (or the *contract multiple*), which is set by the exchange (e.g. for the S&P 500, $z = \$250$). For the S&P 500 futures the smallest price change, a *tick*, is 0.1 index points, representing \$25 in value, while a one-point change in the S&P 500 futures index implies a change in the face value of the futures contract of \$250. If the futures index on the S&P 500 is $F_0 = 1,500$, then the *face value of one futures contract* is \$375,000 (i.e. $FVF = zF_0$).

I.B.3.2.2 Index arbitrage and program trading

In Chapter I.A.7 we used the ‘carrying charge’ approach to derive the no-arbitrage condition for the futures prices on a stock,² which has the following equivalent forms:

$$F = S(1 + (r - \delta)T) \quad (\text{using simple rates})$$

$$F = S(1 + (r - \delta)/q)^{qT} \quad (\text{using discrete compound rates})$$

$$F = S e^{(r^c - \delta^c)T} \quad (\text{using continuously compounded rates})$$

where

T is the number of years (or fraction of a year)

r^c is the (continuously compounded) interest rate (expressed as an annual rate)

δ^c is the (continuously compounded) dividend yield

q is the number of periods per year for compounding (e.g. $q = 4$ for quarterly).

If the equalities above do not hold, then index arbitrage can result in (almost) riskless profits. In large financial institutions real-time data on r , δ , T , S and F are fed into computers and when a violation is detected the arbitrage strategy is applied. Borrowing or lending cash (at r) to undertake the arbitrage transaction is usually done via the repo market.

² See also equation (I.A.7.11).

Example I.B.3.1

Suppose the current stock price S is 200, r^c is 0.10, δ^c is 0.03 and $T = 1/3$ (i.e. four months), then the ‘synthetic’ or ‘fair’ futures price is $F = S e^{(r^c - \delta^c)T} = 200e^{(0.10 - 0.03)0.3333} = 204.72$. However, if the quoted futures price is $F = 210$, then an arbitrage opportunity (i.e. ‘sell high, buy low’) is possible. You sell the futures contract today at $F = 210$, borrow $S = 200$ at $r^c = 0.10$ for four months ($T = 0.3333$) and purchase the underlying stocks today. After receiving dividend payments at the rate $\delta^c = 0.03$ you owe 204.72 after four months. You deliver the stocks against the short futures position after four months and receive $F = 210$ (from the long in the futures), making an arbitrage profit of 5.28 ($= 210 - 204.72$). Conversely if the quoted futures price had been $F = 202$, which is less than $SF = 204.72$, the arbitrage strategy would have involved buying the futures contract ‘low’ at 202 and simultaneously *short-selling* the stocks at $t = 0$ for $S = 200$ and investing the proceeds at the risk-free rate. At expiry of the long futures position you would pay $F = 202$ and receive the stocks, which would be returned to the broker you had borrowed them from at $t = 0$ for your short sales. You would also have had to pay the broker any dividends due on the borrowed stocks, hence your next receipts from investing the $S = 200$ at the risk-free rate would be $S e^{(r^c - \delta^c)T} = 204.72$. Hence your arbitrage profit would be 2.72 (receipts of 204.72 minus payments of 202).

There is some *basis risk* in the arbitrage strategy because the traders only buy (or sell) a proportion of the stocks in the S&P 500, say, so their ‘cash market position’ might not exactly mirror that in the futures index. There are transaction costs to consider (of the order of 0.6% of the value of stocks in the deal, for market makers) and the timing of dividend payments may be uncertain. Also, it might not be possible to do the trades in the futures and the stock market at exactly the same time and in the meantime the prices of some of the underlying stocks might change. If the arbitrage involved \$2m in stocks and $S = 1400$ (say), then with the index multiple being \$250, the number of futures contracts required (see below) would be $N_f = 5.71$ ($= \$2m / (1400 \times \$250)$) and the arbitrage would not be perfect because you cannot purchase fractional contracts. At the expiry date, the institutions unwind their positions and may, for example, sell their stocks using on-the-close orders (i.e. at closing prices) via the NYSE computerised order-processing system known as DOT (‘designated order turnaround’). Because the arbitrage opportunities are calculated with real-time data using a computer, it is also known as *program trading* and is the subject of some controversy when the markets move rapidly.

Although program trading can be tricky, a large number of financial institutions are engaged in this activity. The NYSE defines a program trade as one involving a (near) simultaneous purchase or sale of more than 15 stocks with a market value in excess of \$1m, and all such trades, as well as index arbitrage trades, must be reported to it and a summary is published in the *Wall Street*

Journal. Academic studies suggest that index arbitrage profits are difficult to obtain but occur more often in those contracts with a longer time to maturity. This may be because these are the less liquid contracts, where the fair price and actual price might diverge because traders cannot obtain credit lines for these longer periods and therefore cannot instigate the arbitrage transaction.

I.B.3.2.3 Hedging Using Stock Index Futures

A well-diversified portfolio eliminates unsystematic (or idiosyncratic or diversifiable) risk, but with stock index futures it is possible to hedge against systematic risk. If our hedged portfolio entirely eliminates systematic risk, we might expect it to earn the riskless rate of return (since it is effectively a ‘safe portfolio’). Hedging relies on the positive correlation between spot and futures prices. If you are long in the spot asset, then shorting the futures contract ensures a negative correlation in the hedged portfolio.

A portfolio manager may wish to hedge her stock portfolio if she feels that the cash market is likely to be particularly turbulent over short periods or if she thinks she has ‘winners’ in her portfolio but is worried about a *general fall* in the index. For example, she can hedge her overall position using index futures while still holding some ‘undervalued’ stocks. She therefore isolates the market timing issue from the stock selection problem. She could of course sell all or some of her stocks if she believes their prices will fall, but a more cost-effective method of hedging is to use stock index futures.

Since she is long the underlying stock portfolio, she takes a *short futures hedge* (i.e. she sells index futures). An effective hedge position requires a calculation of the number of futures contracts she needs to short. This can be done in two ways, a ‘simple’ and an ‘optimal’ method, but luckily both tend to give similar hedge outcomes. The simplest way to consider this issue is to assume a spot (or ‘cash’) position of \$1.4m in a portfolio of stocks that exactly mirrors the composition of the S&P 500 index. Hence we are assuming perfect correlation between the index (S) and the futures index (F) (i.e. the correlation coefficient is +1) and both indices move by the same (absolute) amount ($\Delta S = \Delta F$).

If the quoted futures price is $F_0 = 1500$ index points, then the face value of one futures contract is:

$$FVF_0 = zF_0 = \$250 \times 1500 = \$375,000$$

It seems intuitively obvious that the required number of futures contracts should be short 3.73 contracts:

$$N_f = -\frac{\text{Value of Spot Position}}{\text{Face Value of one Futures Contract}} = -\frac{TVS_0}{zF_0} = -\frac{\$1.4\text{m}}{\$250 \times 1500} = -3.73$$

If the investor holds a diversified portfolio but one that moves more than or less than the S&P 500, then the beta β_p of her portfolio is *not* unity. For example, if her portfolio has a beta that exceeds unity then she must sell more futures contracts because the value of her equity portfolio will move more than the futures index. Hence the simple formula above must be amended:

$$N_f = -\frac{\text{Value of Spot Position}}{\text{Face Value of one Futures Contract}} \beta_p = -\frac{TVS_0}{FVF_0} \beta_p$$

The above formula is known as the *minimum variance hedge ratio*.

Let us briefly consider an alternative intuitive argument for calculating N_f^* . Suppose the stock index $S_0 = 1400$. Then *the number of units of the index* held in stocks is:

$$N_{S,0} = -\frac{\text{Value of Spot Position}}{\text{Value of Spot Index at } t = 0} = \frac{TVS_0}{S_0} = \frac{\$1.4\text{m}}{1400} = 1000 \text{ index units}$$

This means that for each $\Delta S = +1$ unit, the stock portfolio changes in value by \$1000 or, alternatively, that the \$1.4m is allocated over 1000 ‘shares’ of the S&P 500. If the index S and the futures on the index F both move perfectly together, it seems logical to hedge the position in the stocks by shorting enough futures contracts to be equivalent to 1000 units of the index. Put slightly differently, the appropriate number of futures contracts is one for each ‘share’ held in the S&P 500. This gives $N_{f,0} = -N_{S,0} = -1000$ index units to be held in futures, which implies you should short four futures contracts:

$$N_f = -N_{S,0} / z \beta_p = -1000 / 250 = -4,$$

since here the composition of our stock portfolio mirrors the S&P 500 and $\beta_p = 1$. The difference between the two methods for calculating N_f involves the denominator of the expression: the ‘simple’ method uses the ‘current value of the futures index’ ($FVF_0 = zF_0$) and the ‘optimal’ method uses ‘the current notional value of the index itself’ (zS_0). If F_0 and S_0 are quite close, the two methods give very similar results. We are now ready to calculate N_f^* for a specific portfolio of stocks held by an investment manager.

In the following example we suppose it is the 20th of June and we wish to hedge a £1m portfolio of UK stocks with $\beta_p = 1.5$ over the next five months. We will have to use the December FTSE 100 futures contract since the hedge period goes beyond the maturity date of the October contract. We find that if there is no basis risk, then shorting 32–33 December futures contracts

will ensure that any change in value of our UK equity portfolio will be largely offset by the change in value of our futures position (see example I.B.3.2).

Example I.B.3.2: Hedging Using Stock Index Futures (FTSE 100)

Data: *It is the 20th of June and a pension fund wishes to hedge its £1m stock portfolio. The hedge period is five months. The beta of the equity portfolio is 1.5. The dividend yield on the equity portfolio is $\delta = 4\%$ and $r = 10\%$ (continuously compounded rates). You choose to hedge with the December futures with $T = 6/12 = 0.5$ years to expiry. The current level of the FTSE 100 is 4500.*

20th June:

$$S_0 = 4500 \text{ (FTSE 100)}$$

$$F_0 = 4500e^{(0.10 - 0.04)/2} = 4637 \text{ (December delivery)}$$

Futures Contract:

Contract multiple = £10 per index point

Tick Size = 0.5

Tick Value = £5

Question:

How many futures contracts (N_f) should you short in order to hedge the equity portfolio?

Answer:

The minimum variance hedge ratio is:

$$N_f = -(TVS/FVF) \times \beta_p = -(\text{£}1\text{m}/(\text{£}10 \times 4637)) \times 1.5 = -32.35$$

If the ‘alternative method’ had been used then:

$$N_{S,0} = TVS_0/S_0 = 1,000,000/4500 = 222.2$$

$$N_f = -N_{S,0} / \beta_p = -222.2/10 \times 1.5 = -33.3$$

Both methods suggest shorting 32–33 futures contracts. Both methods give the same result because, for index futures that have a relatively short period to maturity, F_0/S_0 is quite close to unity. In this case F_0 was 3% larger than S_0 .

The investor can earn the risk-free rate by simply selling the stock portfolio and purchasing T-bills. However, if she thinks the volatility in the stock market will be for only a short period, she will save on transaction costs by using index futures. Also, if she holds some stocks that she believes will do better than average, she can use index futures to offset the systematic risk of the whole portfolio, while reaping the benefits of her ‘stock-picking’ strategy. The risks in hedging with index futures arise if:

- the beta of the stock portfolio is estimated incorrectly or it changes over the hedge period;
- the stock portfolio is not well diversified so that some non-systematic risk is present (and hence the change in value of the stock portfolio is not correctly measured by its beta);

- the safe rate r (or the dividend yield δ) alters over the hedge period (which alters the relationship between F_1 and S_1) – this is basis risk.

The first point is perhaps the most serious and may arise if beta is really time-varying but one has used regression analysis, which assumes a fixed value for beta. The variability in beta can be assessed using advanced regression techniques: similarly, advanced methods for forecasting beta over the hedge period may be developed, but these are beyond the scope of the PRM syllabus. Alternatively, one can note that beta is defined as the ratio of the covariance between the return on the underlying portfolio of stocks and the ‘futures return’ $\Delta F/F$, to the variance of the futures return. Recently, the modelling of time-varying variances and covariances using ARCH and GARCH models (see Cuthbertson and Nitzsche 2001b and Cuthbertson and Nitzsche 2004) has been rather successful especially with ‘high-frequency’ daily data, and these models can be used to predict β . There is some evidence that this may provide some improvement in the hedging outcome compared with using a fixed value for β (see, for example, Baillie and Myers 1991).

Finally, speculation with stock index futures is straightforward since arbitrageurs ensure F and S are nearly perfectly positively correlated. Hence, if you forecast the S&P 500 to rise in the future you can go long (i.e. buy) futures contracts on the S&P 500 – and vice versa. If your forecast is correct you can close out, receiving a profit (from the clearing house) of $N_f z_t (F_1 - F_0)$.

Speculation with futures (rather than buying the stocks in the S&P 500 index itself) provides *leverage*, since you do not have to pay out actual cash for the futures contract at $t = 0$ (you do have to pay a small initial margin payment, but this does not alter the key element in the argument – see Chapter I.C.6).

I.B.3.2.4 Tailing the Hedge

When we derived the formula for the minimum-variance hedge ratio N_f no account was taken of the fact that interest is earned on positive inflows into the margin account and funds must be borrowed if there are margin calls. *Tailing the hedge* makes an adjustment to N_f to take account of these interest payments (receipts), so that the change in the futures position more closely matches the change in the underlying spot (cash) market position. To illustrate, assume the hedger can borrow or lend at the risk-free rate (which is constant) and the hedger goes long one futures contract at F_0 , three days before maturity (and hence has a short position in the underlying asset S). Ignoring interest payments on the margin account, the change in the futures price over the three days is:

$$(F_1 - F_0) + (F_2 - F_1) + (F_3 - F_2) = F_3 - F_0$$

and (if basis risk is small) this offsets any change in the price of the spot asset over three days, $S_3 - S_0$. However, the futures contract is marked to market each day and interest is earned (charged) on any positive (negative) changes in the margin account, so the change in the value of the futures position will not (in general) equal $F_3 - F_0$. For example, assume the futures price rises (monotonically) from $t = 0$ to maturity T at $t = 3$. At the end of the first day a profit of $F_1 - F_0$ will accrue to the long futures position, which will be worth $(F_1 - F_0)e^{r(2/365)}$ at maturity (using continuously compounded rates). But at $t = 0$, with three days to maturity if you *reduced* your (minimum variance) hedge ratio to $N_{f,T-3}^* = N_f / e^{r(2/365)}$, then this would just offset the interest earned on the variation margin of your long position. Similarly, if you continue to tail the hedge with two days to maturity, the number of long futures contracts held would be $N_{f,T-2}^* = N_f / e^{r(1/365)}$, and with one day to maturity you would hold N_f contracts (since you close out and no interest accrues on the final day). In practice, traders would tail the hedge at the outset of the contract but would adjust this figure only somewhat infrequently, in order to save on transaction costs (and the optimal rebalancing for tailing the hedge would be calculated by computer software). Note that at $t = 0$, with n days to maturity, tailing the hedge requires a position in:

$$N_{f,T-n}^* = N_f / e^{r((n-1)/365)} \approx N_f [1 - r((n-1)/365)]$$

for $|r(n-1)/365| \ll 1$. Hence if you have calculated your minimum-variance hedge ratio N_f , the number of contracts to tail the hedge is (approximately) $N_{\text{tail}} = N_f (n-1)(r/365)$. So, if the hedge involves a long position in the futures $N_f > 0$, *tailing the hedge* requires you to short a (small) number of contracts or, in other words, to reduce the number of long contracts given by the minimum-variance hedge ratio. It should be obvious that the adjustment required to tail the hedge becomes important only when the time to maturity is relatively long (i.e. more than a few months). For example, for a hedge period of $n = 180$ days to maturity, $r = 3\%$ p.a. and $N_f = 100$ contracts, then $N_{\text{tail}} = -1.48$ contracts, which is not a huge adjustment. In the remainder of this chapter we calculate hedge ratios ignoring the problem of tailing the hedge.

I.B.3.2.5 Summary

- Index futures prices can be determined using the cost-of-carry approach, and violations of this relationship give rise to index arbitrage (or program trading).
- Stock index futures provide a low-cost method of hedging an existing diversified equity portfolio or of hedging a future purchase of a diversified stock portfolio (i.e. an anticipatory hedge).
- Hedging is not perfect because (i) the equity portfolio held will in general not exactly replicate the composition of the index future and (ii) purchases or sales of the underlying

stocks cannot be perfectly synchronised with transactions in the futures market. There are also transaction costs to consider and the fact that the portfolio beta is an estimate and may not be accurate over the hedge period. These factors give rise to basis risk in the hedge.

- Speculation with stock index futures provides leverage since you do not have to ‘pay upfront’ for the futures contract (as you do in the spot (cash) market). If you believe stock prices will rise (fall) then you would go long (short) the futures contract – and close out at a later date, hopefully at a profit!

I.B.3.3 Currency Forwards and Futures

The spot rate is the exchange rate quoted for immediate delivery of the currency to the buyer (actually, delivery is two working days later). A second type of deal involves the forward rate, which is the guaranteed price agreed today at which the buyer will take delivery of currency at some future period. Currency forwards and futures are very similar analytically, even though in practice the contractual arrangements differ between the two types of contract.

Both a currency forward contract and a currency futures contract are obligations to trade one currency for another at a specified exchange rate on a specific future delivery date. The dealing costs in both type of contract are very small relative to the size of the average deal. Why, then, do we need both types of contract? The reasons are the slight variations between the two types of contract.

A currency forward deal can be designed to exactly fit the client’s requirements as to the principal amount in the trade, its exact delivery date and which currencies are involved. It is an OTC contract. Most of the forward deals on a very wide range of currencies are channelled through London, New York and Tokyo, with Hong Kong and Singapore being influential in the Far East. This OTC market is very efficient, with low transaction cost.

Deals are negotiated between large multinational banks (e.g. Merrill Lynch, Citicorp, Morgan Stanley) for delivery of one currency (e.g. Swiss francs) usually against the US dollar (the vehicle currency) at a specific date in the future. It is also possible to negotiate so-called forward cross-rate deals (which do not involve the US dollar), for example, delivery of Swiss francs and the receipt of euros. For most major currencies, highly liquid markets are for one- to six-month maturities and, in exceptional circumstances, three to five years ahead. Use of the forward market eliminates any risk from subsequent changes in spot rates. This is because the forward rate is agreed today, even though the cash transaction takes place in (say) one year’s time.

In contrast, a currency future is an exchange-traded instrument on a limited set of currencies. The contract size is fixed (e.g. SFr 125,000), as are the set of delivery dates. With a forward contract, delivery usually takes place, whereas with a futures contract you do not necessarily have to take delivery of the currency because you can easily close out your position. Finally, the futures contract is marked to market and has little or no credit risk, whereas the forward contract does involve credit counterparty risk.

Currency forwards and futures can be used to hedge future receipts or disbursements in foreign currency by, for example, exporters and importers, or receipts or disbursements of capital assets such as foreign bonds, stocks, bank deposits and bank loans. They can also be used to provide leverage in speculative transactions between currencies.

There are a large number of currency futures traded on different exchanges, the major one being the IMM. For example, on the IMM the following currencies are traded against the US dollar: pound sterling, euro, yen, Swiss franc, Canadian dollar, Australian dollar and Mexican peso, as well as other currencies. Other notable centres trading FX futures are the Singapore International Money Exchange (SIMEX) and the Sydney Futures Exchange (SFE).

I.B.3.3.1 Currency Forward Contracts

The pricing of a forward contract involves a relationship between the forward rate and three other variables, the spot rate and the money market interest rates in the two countries, and this is known as *covered interest parity* (CIP). We shall see that (in an efficient market) the quoted forward rate ensures that no riskless arbitrage profits can be made by transactions in the spot currency market, the two money markets and the forward market. CIP is an equilibrium ‘no-arbitrage’ condition. The CIP relationship between spot and forward rates can be derived as follows.

Example I.B.3.3

Assume that a UK corporate treasurer has a sum of money, £ A , which she can invest in the UK or the USA for one year, at which time the returns must be paid in sterling. We assume the forward transaction has no credit risk. Therefore, for the treasurer to be indifferent as to where the money is invested, it has to be the case that returns from investing in the UK equal the returns in sterling from investing in the USA. Assume a UK corporate treasurer is faced with the following interest rates in the ‘domestic’ money market (sterling) and in the ‘foreign’ money market (US dollar). The spot exchange rate (S) is measured as the domestic per unit of foreign currency (i.e. £/\$) as is the one-year quoted forward rate F .

$$\begin{array}{ll} r_d = 0.11 \text{ (11 \%)} & r_f = 0.10 \text{ (10 \%)} \\ S = 0.6667 \text{ (£/\$)} & \text{(equivalent to 1.5 \$/£)} \end{array}$$

$$F = 0.6727 (\text{£}/\text{\$}) \quad (\text{equivalent to } 1.4865 \text{\$/£})$$

We can show that these figures imply that equal returns will result from investing in either the UK or the USA.

1) Invest in the UK

$$\text{In one year receive (terminal value) } TV_{UK} = \text{£}A(1 + r_d) = \text{£}100(1.11) = \text{£}111$$

2) Invest in the USA

a) Convert $\text{£}100$ to $\text{\$}(100 / 0.6667 (\text{£}/\text{\$})) = \text{\$}150$ in the spot market *today*, then

b) Invest in dollar deposits, and dollar receipts at end year are:

$$\text{\$}(100 / 0.6667 (\text{£}/\text{\$})) \times 1.10 = \text{\$} (A/S) (1 + r_f) = \text{\$}165$$

c) Enter into a forward contract *today* for delivery of sterling in one year's time and be certain of receiving a terminal value TV (in £ s):

$$TV_{us} = \text{£}[(A/S)(1 + r_f)]F = \text{£}[(100 / 0.6667 (\text{£}/\text{\$})) \times 1.10] 0.6727 (\text{£}/\text{\$}) = \text{£}111$$

All of the above transactions (a)–(c) are undertaken today at known ‘prices’ – therefore there is no (market) risk. The UK corporate treasurer will be indifferent in placing her funds in the USA or the UK – this ‘no-profit’ condition is the CIP relationship.

Since both investment strategies are riskless arbitrage will ensure that they give the same terminal value: $TV_{UK} = TV_{us}$, or $\text{£}A(1 + r_d) = \text{£}[(A/S)(1 + r_f)]F$. Hence CIP can be expressed as:

$$F/S = (1 + r_d)/(1 + r_f) \quad (\text{I.B.3.1})$$

Or, subtracting ‘1’ from each side:

$$(F - S)/S = (r_d - r_f)/(1 + r_f)$$

The above formulae are exact, but, if r_f is small (e.g. 0.03), then $1 + r_f$ is approximately equal to one and the CIP condition approximates to:

$$(F - S)/S \approx r_d - r_f$$

that is,

$$\text{forward premium (discount)} \approx \text{interest rate differential}$$

It is easy to check that the above data are consistent with this condition:

$$\text{Interest differential} = (r_d - r_f)/(1 + r_f) = (0.11 - 0.10)/1.10 = 0.0091 \text{ (or } 0.91\%)$$

$$\text{Forward discount on sterling} = (F - S)/S = 0.91\%$$

The CIP condition is an equilibrium condition based on riskless arbitrage. If CIP does not hold, there are forces that will quickly restore equilibrium. For example, if $F = S$ but $r_d(\text{UK}) > r_f(\text{US})$, then US (foreign) residents would want to earn a risk-free profit by purchasing UK bills, pushing their price up and interest rates down. (Alternatively, they would all want to invest in Eurosterling deposits and the banks would lower their deposit rates.) To purchase UK assets, US residents would also have to buy spot sterling, and to cover their position they would have to simultaneously sell dollars forward. Hence spot sterling would appreciate (i.e. S would fall) and F would rise. These changes in r , S and F will tend to quickly restore the parity relationship.

One further ‘trick’ to note is that the CIP formula looks slightly different if S and F are measured as ‘foreign per unit of domestic currency’. However, the following ‘rule of thumb’ always holds: If S (or F) is measured as ‘domestic currency per unit of foreign currency’ (i.e. domestic/foreign), then the interest rates are also domestic/foreign in equation (I.B.3.1). For example, if S and F are measured as Swiss Francs per US dollar then the CIP condition is:

$$F/S = (1 + r_{SF})/(1 + r_{\$})$$

where r_{SF} and $r_{\$}$ are the interest rates in Switzerland and the USA, respectively. Conversely, if S and F are measured as US dollars per Swiss franc, then the US interest rate would appear in the numerator and the Swiss Franc interest rate in the denominator of the above formula.

I.B.3.3.2 Currency Futures Contracts

We now switch from forward contracts to futures. The most actively traded futures contracts traded on the CME are in the euro, Canadian dollar, Japanese yen, Swiss franc and sterling (see Chapter I.C.6). On the IMM there is ‘after-hours’ electronic futures trading using the Globex system and euro futures can be traded at any time if one is willing to trade on different exchanges. We will concentrate on the euro futures contract on the CME. If you are long one contract then the convention is: *Long Euro Futures* \Rightarrow *Receive €125,000 and pay out US dollars.*

Example I.B.3.4

Suppose Ms A on 1st March purchases (i.e. is long in) one euro September futures contract at a futures rate of $F_0 = 1.04(\$/\text{€})$. At expiry (in September) she would *receive* €125,000 and pay out \$130,000. The contract is marked to market. So, for example, if, on 2nd March, $F_1 = 1.0450(\$/\text{€})$, then Ms A has gained since (forward) euros have appreciated and her margin account will be credited with \$625 (i.e. 50 ticks \times \$12.50) – see also Chapter I.C.6.

I.B.3.3.3 Hedging Currency Futures and Forwards

International investors and multinational corporations are vulnerable to *transactions exposure*, namely exchange-rate risk of future cash receipts or payments in a foreign currency. Clearly, forward contracts (which are OTC agreements) that are held to maturity can be used to hedge FX risk by ‘locking in’ the rate of exchange that will apply at some future date.

Currency futures can also be used to hedge this exposure and because the correlation between spot and futures prices is high the hedging error is small, even though you close out the futures contract some time before maturity. If the investor is long in the spot market (i.e. holds the foreign currency) she will short futures contracts (and vice versa).

Example I.B.3.5

Let us consider a US importer (on 1st April) who has to pay SFr 500,000 in six months’ time (on 1st October) for imports from Switzerland. If the Swiss franc appreciates (i.e. the dollar depreciates) over the next six months then she will have to pay out more dollars (than at the current spot rate). To hedge this position she takes a long position in Swiss franc futures. What she loses in the spot market she hopes to offset with gains in the futures market (and vice versa), when she closes out the futures. The hedge position is: *Required to pay SFr in six months* \Rightarrow *Buy SFr Futures today*. Suppose she is currently faced with the following rates:

$$S_0 = \text{spot rate} = 0.6700(\$/\text{SFr})$$

$$F_0 = \text{futures price (Oct. delivery)} = 0.6738(\$/\text{SFr})$$

$$\text{Contract size, } \mathcal{Z} = \text{SFr } 125,000$$

$$\text{Tick size (value)} = 0.0001(\$/\text{SFr}) = \$12.50$$

Naïve hedge ratio (1st April, $t = 0$): First we calculate the number of futures contracts required to cover the spot position. This can be done using either the Swiss franc or the US dollar as the ‘common currency’, but the two methods usually give similar answers. (Which is more useful depends in part on whether we expect the absolute change in F and S or their proportionate changes to be most closely correlated.)

Method 1: The simplest method is to use the SFr amount outstanding, $TVS_0 = \text{SFr } 500,000$, and divide by the contract size, $\mathcal{Z} = \text{SFr } 125,000$. This is usually accurate enough and we will concentrate on this method in the examples in this chapter.

$$N_f = \frac{\text{Cash Market Position in SFr}}{\text{Contract Size}} = \frac{TVS_0}{\mathcal{Z}} = 4 \text{ contracts}$$

Method 2: Slightly more complex is to convert both the cash market position and the size of the futures contract into US dollars, the currency of the US importer, then:

$$N_f = \frac{\$ \text{ value of spot position}}{\$ \text{ value of one futures contract}} = \frac{\$TVS_0}{\$FVF_0}$$

$$= \frac{(TVS_0)S_0}{zF_0} = \frac{500,000 \times 0.6700}{125,000 \times 0.6738} = \frac{\$335,000}{\$84,225} = 3.98 \quad (4 \text{ contracts})$$

where TVS_0 is the total value of spot exposure in Swiss francs

$\$TVS_0$ is the total value of spot exposure in US dollars

S is the spot (\$/SFr) exchange rate

F is the price of Swiss franc futures

z is the size of contract (SFr 125,000)

$\$FVF_0$ is the face value in US dollars of one Swiss franc futures contract ($= zF_0$).

We can analyse hedging by the US importer either in terms the *change* in the spot position net of the *change* in the futures position or in terms of the *dollar value* of the payments made by the US importer. We discuss each in turn. In Example I.B.3.6 we show the position on 1st October (at $t = 1$) after the Swiss franc has appreciated from $S_0 = 0.67(\$/\text{SFr})$ to $S_1 = 0.72(\$/\text{SFr})$ (i.e. 500 ticks). The increase in cost to the US importer in the spot market is $TVS_0(S_1 - S_0) = \$25,000$. However, as the Swiss franc spot rate appreciates, so does the futures price from $F_0 = 0.6738(\$/\text{SFr})$ to $F_1 = 0.7204(\$/\text{SFr})$, a change of 466 ticks. The gain on the long futures position is \$23,300 ($= 466 \text{ ticks} \times \$12.50 \times 4 \text{ contracts}$, which can also be expressed as $N_f z (F_1 - F_0) = \$23,300$). The gain on the futures position of \$23,300 nearly offsets the increased cost in the spot market of \$25,000.

Example I.B.3.6: Long hedge (Swiss franc futures)

On 1st of April

$S_0 = 0.6700(\$/\text{SFr})$

US importer has to pay

$TVS_0 = \text{SFr } 500,000$ on 1 October

$\$TVS_0 = \text{Dollar value of imports} = (TVS_0)S_0 = \$335,000$

$z = \text{Contract size} = \text{SFr } 125,000$

$F_0 = 0.6738(\$/\text{SFr})$

$FVF_0 = F_0 z = \$84,225$

$$N_f = \frac{\$ \text{ value of spot position}}{\$ \text{ value of futures contract}} = \frac{TVS_0}{FVF_0} = \frac{335,000}{84,225} = 3.98 \quad (4 \text{ contracts})$$

Aide memoire: Initial basis $= b_0 = (S_0 - F_0) = -0.0038(\$/\text{SFr}) = -38 \text{ ticks}$

On 1st of October

$S_1 = 0.7200(\$/\text{SFr})$ (SFr has appreciated, \$ has depreciated).

$F_1 = 0.7204(\$/\text{SFr})$

US importer pays spot

$$\$TVS_1 = (TVS_0)S_1 = \text{SFr } 500,000 \times 0.7200 = \$360,000$$

$$\text{Extra \$ spot payments} = \$TVS_1 - \$TVS_0 = TVS_0 (S_1 - S_0) = \$25,000$$

$$\begin{aligned} \text{Gain on futures} &= N_f \zeta (F_1 - F_0) = 4(125,000) (0.7204 - 0.6738) = \$23,300 \\ &= 4 (\$12.50) (466 \text{ ticks}) = \$23,300 \end{aligned}$$

Aide memoire: Final basis, $b_1 = S_1 - F_1 = -0.0004$ (4 ticks)

Change in basis, $\Delta S - \Delta F = b_1 - b_0 = (-4) - (-38) = +34$ ticks

Analysis:

The rise in spot Swiss francs makes the imports more expensive, but this is offset by the profit on the four futures contracts. One way of looking at the hedge outcome is that the gain on the futures of \$23,300 is about equal to the loss in the cash market of \$25,000.

Alternative Method:

$$\begin{aligned} \text{Net \$ cost of imports} &= \text{Cost in spot market} - \text{Gain on futures} \\ &= TVS_0 S_1 - N_f \zeta (F_1 - F_0) = \$ 360,000 - \$ 23,300 = \$336,700 \end{aligned}$$

Let us now examine the hedge from the slightly different objective of the final dollar cost to the US importer. In October the net cost is:

$$\begin{aligned} \text{Net \$ cost of imports} &= \text{Cost in spot market} - \text{Gain on futures} \\ &= TVS_0 S_1 - N_f \zeta (F_1 - F_0) \\ &= TVS_0 (S_1 - F_1 + F_0) = TVS_0 (b_1 + F_0) \\ &= \$ 360,000 - \$ 23,300 = \$ 336,700 \end{aligned}$$

Note that we have used $N_f \zeta = TVS_0$. The calculations above make clear that the hedge ‘locks in’ the futures price at $t = 0$ (i.e. F_0) as long as the final basis $b_1 = S_1 - F_1$ is ‘small’. Effectively, the importer pays out \$336,700 to receive SFr 500,000, which implies an effective rate of exchange at $t = 1$ of:

$$\frac{\text{Net Cost}}{TVS_0} = \frac{\$ 336,700}{\text{SFr } 500,000} = b_1 + F_0 = 0.6734 (\$/\text{SFr}),$$

which is very close to the initial futures price at $t = 0$ of $F_0 = 0.6738(\$/\text{SFr})$, the difference being the final basis $b_1 = -4$ ticks.

The effective cost of the imports is close to that which would have occurred had the US importer used a *forward contract* on 1st April at a cost of \$ 336,900 ($= F_0 TVS_0 = 0.6738 \times 500,000$). Since for the naïve hedge ratio $TVS_0 = N_f \zeta$, the net cost of the imports $= TVS_0[S_1 - (F_1 - F_0)] = TVS_0[b_1 + F_0]$, hence F_0 is ‘locked in’ provided the final basis is small.

Example I.B.3.7

An example of a short hedge using futures contracts requires the US resident to have a long position in the foreign spot market. The latter might arise if on the 1st of April a US multinational expects to receive sterling payments on the 1st of October either from sales of goods in the UK or from sterling investments. Hence the US multinational is ‘long sterling’ in the spot market and fears a fall in sterling, hence it would hedge by shorting sterling futures contracts: *Will receive sterling in 6 months* \Rightarrow *Sell Sterling Futures today*. If spot sterling depreciates between April and October the US multinational loses in the spot market. However, it will gain by closing out the short sterling futures position at a profit, since a fall in spot sterling will usually be accompanied by a fall in the futures price. Of course, these hedging strategies involve basis risk: if there are sharp changes in either US interest rates or those of the foreign country then F and S will not move together and the hedge may perform much worse than expected.

I.B.3.3.4 Summary

- Currency forwards are OTC contracts, which can be designed to exactly fit the client’s requirements as to amount, delivery dates and currencies. Currency forwards can be used to ‘lock in’ a known exchange rate and hence can be used to hedge foreign currency receipts or payments.
- Currency forwards and futures prices are determined by a no-arbitrage condition known as covered interest parity.
- Riskless arbitrage then ensures that the quoted forward rate $F = S(1 + r_f)/(1 + r_d)$, where F and S are measured as ‘domestic currency per unit of foreign currency’.
- Currency futures are similar to currency forwards, the main difference being that futures can be easily closed out prior to the maturity (delivery) date of the contract. Currency futures provide a low-cost method of hedging known future foreign currency receipts or payments (i.e. transactions exposure). There is virtually no credit risk because of the use of margin requirements.

I.B.3.4 Commodity Futures

Futures on many commodities such as grains and oilseed (e.g. wheat, soybeans, sunflower oil), food (e.g. cocoa, orange juice), metals and petroleum (e.g. gold, silver, platinum, heating oil), and livestock and meat (e.g. live hogs, live cattle, pork bellies) are actively traded on various exchanges around the world (e.g. New York Cotton Exchange (NYCE), Kansas City Board of Trade (KCBT), Mid America Commodity Exchange (MCE), New York Mercantile Exchange (NYMEX) and London Metals Exchange (LME)).

These contracts can be priced using cash-and-carry arbitrage, but we have to take account of storage costs and the so-called convenience yield. If γ is the proportionate (per-period) storage cost, then buying the spot asset at $t = 0$ for S , would result in a cost $S(1 + \gamma T)(1 + rT)$ at maturity of the futures at T (see equation (I.A.7.10)). Hence arbitrage would ensure that the futures price is always higher than the spot price, as shown below:

$$F = S(1 + \gamma T)(1 + rT) = S + \chi \quad (\text{I.B.3.2})$$

where the (dollar) cost of carry is the positive value $\chi = S(r + \gamma)T$ if we ignore the term $(S\gamma)T^2$, which is ‘small’. However, it is often observed that commodity futures are in *backwardation* (i.e. $F < S$). This outcome is rationalised by considering the *convenience yield* v which is defined as:

$$v \equiv F - (S + \chi)$$

The convenience yield arises because the holder of a spot commodity (e.g. soybeans) has the added advantage that she can supply her customers if the commodity goes into short supply (e.g. just before the soybean harvest) and hence retain customer loyalty. This ‘convenient’ state of affairs has a value, which is referred to as the convenience yield.

Note that the convenience yield has nothing to do with speculation about the future path of spot prices. The presence of a convenience yield might therefore prevent cash-and-carry arbitrage operating when the current spot price is high (an indication that the commodity is in short supply) relative to the futures price given by equation (I.B.3.2). Since it is difficult to short-sell commodities, this could prevent the fair futures price being exactly determined by arbitrage. For example, if the spot price of crude oil is very high relative to the futures price, then arbitrage requires short-selling spot crude and simultaneously buying the futures. This would require borrowing crude oil to satisfy delivery in the short sale in the cash market. But if tanker owners will not deliver the spot oil to the required destination then such arbitrage is impossible. There is therefore not such a close link between changes in spot and futures prices for commodities as there is for contracts on other spot assets (e.g. stock indices, currencies) and hence hedging becomes more problematic, with basis risk being higher than for other futures contracts.

I.B.3.5 Forward Rate Agreements

A forward rate agreement (FRA) is a form of forward contract that allows one to ‘lock in’ or hedge interest-rate risk over a time-specific period starting in the future. If you are borrowing money in the future (e.g. in three months’ time), you may fear a rise in interest rates since your future borrowing costs will be higher. Conversely, if you are thinking of lending money in the future, you will lose out if interest rates fall, since your deposit will earn less interest. Hence:

To hedge a rise in future borrowing rates \Rightarrow buy ('go long') an FRA

To hedge a fall in future lending rates \Rightarrow sell ('short') an FRA

Example I.B.3.8

Suppose Ms B has borrowed \$1m from a bank for six months with interest payments reset at the end of the first three-month period, based on the prevailing three-month LIBOR. For the first three-month period the rate on the loan is fixed at the *current* three-month LIBOR rate at $t = 0$, and she can do nothing about this. Assume, however, that Ms B thinks LIBOR rates will rise in the future, hence increasing the cost of rolling over the loan three months from now. She can hedge this risk by buying a 3×6 FRA today at an agreed FRA rate $f_{3,6}$ to begin in three months' time and lasting for a further 90 days. She also continues with her original loan from her 'correspondent' bank. The payoff to Ms B from the FRA (assuming 'actual/360' day count) is:

$$\text{Payoff from long FRA at } t + 3 = (\text{LIBOR}_{t+3} - \text{Agreed FRA rate, } f_{3,6}) \times (90/360)$$

Hence, if the LIBOR rate at $t + 3$ months is greater than the fixed rate agreed in the FRA, the seller of the FRA will pay Ms B the difference. (As we shall see below, there is a further practical nuance because this 'payoff' is calculated at $t + 3$ months but does not accrue until a further three months.) If Ms B continues with her bank borrowing but also holds the (long) FRA then the effective cost of the loan repayments in three months' time is:

$$\begin{aligned} \text{Effective cost of loan} &= (\text{LIBOR}_{t+3} - \text{gain from FRA}) \times (90/360) \\ &= [\text{LIBOR}_{t+3} - (\text{LIBOR}_{t+3} - f_{3,6})] \times (90/360) = f_{3,6} \times (90/360) \end{aligned}$$

Hence, by taking out the FRA, Ms B has effectively swapped her floating-rate payments on her original bank loan for fixed-rate payments at the rate $f_{3,6}$ negotiated in the FRA. Of course, if LIBOR rates at $t + 3$ months turn out to be lower than the FRA rate $f_{3,6}$ (negotiated at $t = 0$), Ms B will pay the seller of the FRA the difference. However, the payments on her original bank loan will be at the lower LIBOR rate, so again she effectively ends up with net payments (i.e. the bank loan plus the FRA) equal to the fixed rate $f_{3,6}$ in the FRA.

Banks are the main participants in the FRA market, which are OTC instruments. That is, they are not traded continuously in a standardised format on an exchange but can be tailored to suit individual requirements. By taking a long position in the FRA, Ms B has effectively 'swapped' her payment on her bank borrowing at an unknown floating rate (i.e. LIBOR), for a known fixed-rate payment at the FRA rate $f_{3,6}$. In fact, an interest-rate swap is nothing more than a series of FRAs over a number of reset periods. For example, a five-year swap with interest payments reset every three months is equivalent to 19 ($= 20 - 1$) separate FRAs. (Note that there are 19 FRAs because

the first payment in the swap is at the current known LIBOR rate at $t = 0$. See Chapter I.B.4 and Cuthbertson and Nitzsche (2001b) for further details on swaps.)

I.B.3.5.1 Settlement Procedures

Now let us consider the actual settlement procedure for an FRA. An FRA is a contract on a notional principal amount $\$Q$. However, there is no exchange of principal at the beginning and end of the contract – only the difference in interest payments on the notional principal amount is paid. The settlement (or benchmark) rate used in the contract is usually the LIBOR rate (or the equivalent interbank rate in any given currency).

We have already considered the ‘payoff’ from the FRA and noted that this can be calculated at $t + 3$ months, but the interest applies over the period $t + 3$ months to $t + 6$ months (for a 3×6 FRA). Hence the amount actually paid at $t + 3$ months is the *present value* of the difference between LIBOR at the settlement date and the quoted forward rate agreed at the outset of the contract.

Example I.B.3.9

Consider the payment at settlement on the following 3×6 FRA:

- i) $f_{3,6} = 8\%$ (set at $t = 0$) on a notional principal of $Q = \$1\text{m}$
- ii) Actual three-month LIBOR rate in three months’ time (i.e. at $t + 3$ months), $r_{t+3} = 10\%$

Cash settlement is made at the end of three months. To calculate the cash settlement note that the amount owed at the end of six months is $Q (r_{t+3} - \text{FRA rate}) 90/360$. Since payment takes place after 3 months (i.e. as soon as r_{t+3} becomes known) the amount owing needs to be discounted, hence:

$$\begin{aligned} \text{Actual Payment (after 3m)} &= \frac{1}{1 + 0.10 / 4} [\$1\text{m} (0.1 - 0.08) \times (90 / 360)] = \\ &= \$5000 / 1.025 = \$4878.05 \end{aligned}$$

Only the ‘difference in value’ is paid because Ms B will already have long-term variable-rate loan arrangements with her own bankers and she uses the FRA to hedge *changes* in interest rates. In effect, as we noted earlier, she ‘transforms’ the floating-rate payments on her bank borrowing into a known fixed payment at $f_{3,6}$.

I.B.3.6 Short-Term Interest-Rate Futures

Interest-rate futures became of increasing importance in the late 1970s and early 1980s, when the volatility of interest rates increased. This was partly because of higher inflation and consequent attempts by central banks to control the money supply and exchange rates by altering interest rates. In the USA, Eurodollar futures (CME/IMM) are one of the most actively traded contracts, much more so than Euroyen and US T-bill futures. Three-month sterling futures contracts ('short-sterling') are actively traded on LIFFE. Table I.B.3.2 shows contract specifications for 3-months Sterling interest rate, 90-day Eurodollar and US T-bill futures.

Table I.B.3.2: Contract specifications

Sterling interest rate, Eurodollar and US T-bill futures

	Sterling 3-month (LIFFE)	A. 90-day Eurodollar (IMM) and B. US T-bill (IMM)
Unit of Trading	£500,000	\$1m
Delivery Months	Mar/June/Sept/Dec	Mar/June/Sept/Dec
Quotation	100 – futures interest rate (%p.a.)	Euro \$: IMM index = 100 – futures <i>yield</i> T-bill: IMM index = 100 – futures <i>discount rate</i>
Tick Size (Value)	0.01 (£12.50)	Both = 1 bp (\$25)
Settlement	Cash	Euro \$: Cash T-bill: 90-92 day T-bill
Last Trading Day	3rd Wednesday of delivery month	Euro-\$: 2nd business day before 3rd Wednesday of delivery month T-bills: Business day prior to issue date of 'new' T-bills
First Delivery Day	1st business day after last trading day	Euro-\$: Cash settled on last trading day T-bills: 1st business day of the delivery month

Notes: Margins change frequently. Margin requirements for speculators are higher than for hedgers & spreaders. No daily price limits. Trading hours IMM are 7.20 a.m.–2 p.m. Central Time.

I.B.3.6.1 US T-bill Futures

The underlying asset in the T-bill futures contract is of course a US T-bill, which when issued is usually a 91-day bill, sold at a discount (and is quoted on an 'actual/360'-day basis; see Chapter I.C.2 and Cuthbertson and Nitzsche (2001a)).

Example I.B.3.10

Take a T-bill with 90 days to maturity, face value of \$100 and a quoted *discount rate* d of 8%. This has a market price P of \$98 given by:

$$P = 100 \left[1 - \frac{d}{100} \times \frac{90}{360} \right]$$

This T-bill has a discrete compound yield equal to:

$$y_{TB} = (\$100/\$98)^{365/90} - 1 = 0.08538 \quad (8.538\% \text{ p.a.})$$

Now let us turn to the IMM futures contract on T-bills. This contract allows delivery of 90-, 91- or 92-day T-bills but the futures price is based on the 90-day bill. The contract size is \$1m. It may be somewhat confusing, but this futures contract has a futures discount rate d_f , an index price Q_f and a futures price F , all of which are interrelated.

Example I.B.3.11

If a futures contract has a quoted discount rate $d_f = 8\%$, then the (IMM or CME) quoted index is $Q_f = 92$ ($= 100 - 8$), hence the quoted index price $Q_f = 100 - d_f$.

However, somewhat paradoxically, the IMM index is *not* the actual futures price at which the contract is traded. For example, if the quoted IMM Index is $Q_f = 92$ for the June contract then the actual futures price F (per \$100 nominal) is given by:

$$F = 100 - (100 - Q_f)(90/360) = 100 - (100 - \text{IMM index})/4 \quad (\text{I.B.3.3})$$

Hence, for the June contract with $Q_f = 92$, $F = \$98$ (per \$100). The contract size is $\zeta = \$1\text{m}$ of T-bills and therefore with $F = \$98$ the ‘face value of one futures contract’ or the invoice price is:

$$FVF_0 = \zeta (F_0 / \$100) = \$1\text{m} (\$98 / \$100) = \$980,000$$

Note that the T-bill futures price F is determined in the same way as the price of the underlying spot T-bill, since $F = 100 - (100 - Q_f)(90/360)$ and if we substitute $Q_f = 100 - d_f$ then:

$$F = 100 \left[1 - \frac{d_f}{100} \times \frac{90}{360} \right] \quad (\text{I.B.3.4})$$

Hence for the futures contract, the quoted discount rate d_f of 8% p.a. is equivalent to 2% over 90 days and hence the price of the futures contract (per \$100 nominal) is $F = \$100 - \$2 = \$98$, which is the value given by equations (I.B.3.3) and (I.B.3.4).

The reason why the ‘quoted’ IMM index appears in the *Wall Street Journal* is that it allows traders to quickly assess their gains and losses, since *a one-basis-point change in the IMM Index Q_f (or the discount rate d) corresponds to a \$25 change in the contract price*. The tick size is 1 bp and the tick value is \$25.

For example, if the IMM index falls by 1 bp from 92.00 to 91.99 the new futures price is $F = 97.9975$ and the price of one futures contract (for \$1m nominal) is \$979,975, a fall of \$25. Alternatively, from equation (I.B.3.3), a change of 1 bp (i.e. $\Delta Q_f = 0.01$) gives a change in F of 0.0025 ($= \Delta Q_f / 4$) and hence \$25 per \$1m nominal. This method of price quotation also ensures that the bid price is below the offer (ask) price.

Although 90-day T-bills are usually delivered, the contract also allows delivery of 91- or 92-day T-bills, and if these T-bills are delivered at maturity then 91 or 92 is used in place of ‘90’ in the formula for F , the amount payable. Also, the delivery date on the futures is often timed to coincide with the date on which US 365-day T-bills have between 90 and 92 days to maturity, and this facilitates a liquid spot market for delivery. The contract expiration months are March, June, September and December.

I.B.3.6.2 Three-Month Eurodollar Futures

A Eurodollar deposit (in the cash market) is a three-month (dollar) time deposit in a foreign bank or an overseas branch of a US bank (see Chapter I.C.2). The day count convention is actual/360 and it is quoted on a yield (or ‘add-on’) basis. Hence, if the quoted yield on a 90-Eurodollar deposit is $y = 12\%$ p.a., a deposit of \$100m will accrue to \$103m [$= \$100(1 + 0.12(90/360))$] after 90 days. (This implies an annual compound rate of 12.74% [$= (\$103/\$100)^{365/90} - 1 = 0.1274$].)

Now let us turn to the futures contract for which the underlying asset is a three-month deposit. Because Eurodollar deposits are non-transferable, the futures contract is settled in cash. Settlement on the last trading day is determined by an average of three-month LIBOR rates (as determined by CME officials). The futures contract is written on three-month (90-day) LIBOR (Eurodollar) rate and the futures IMM index quote Q_f in the *Wall Street Journal* can be translated into the true futures price F , as in equation (I.B.3.3) for T-bills (again using the 90/360 basis).

Since each Eurodollar futures contract is for \$1m, a change of 1 bp (1 tick) in the IMM quote Q_f corresponds to a change (tick value) of \$25 in the contract price. The quoted ‘settlement yield’ is simply, $y_f = 100 - Q_f$.

Delivery months are March, June, September and December and extend out for 10 years because these contracts are widely used to hedge either the floating interest-rate payments (or receipts) on long-term bank loans and deposits or a swap dealer's net floating-rate position (see Chapter I.B.4). Note that there is also a Eurodollar futures contract on 30-day deposits.

I.B.3.6.3 Sterling Three-Month Futures

This futures contract is written on a sterling three-month deposit for a notional value of £500,000 per contract. Maturities extend to four years and the nearby contracts are most actively traded. The futures price F is simply:

$$F = (100 - \text{futures interest rate quote \% p.a.}) = 100 - f$$

A change in the *annual* interest rate of 1 bp (0.01% p.a.) is equivalent to 0.25 bp over three months, hence $\Delta F = 0.0025$, which, with a contract size of £500,000, amounts to a change in value of £12.50. *The tick size is 1 bp (0.01% p.a.), which with a contract size of £500,000 gives a tick value of £12.50 on the short sterling (three-month) futures.*

As can be seen from the above equations, all interest-rate futures contracts have the feature that the futures price F and the futures interest rate move in opposite directions. Another important feature is that the futures contracts 'deliver' an asset that reaches maturity sometime later. This is easiest to see for the long T-bill futures, which mature at t_1 and deliver a 90-day T-bill, which matures at $t_2 = t_1 + (90/360)$. For the sterling three-month futures the underlying asset is a *notional* three-month deposit, delivered at the expiration of the futures at t_1 (and which notionally matures at t_2).

I.B.3.6.4 Hedging Interest-Rate Futures

The following illustrates how interest-rate futures allow investors to hedge spot positions in money market assets, such as T-bills and (Eurodollar) deposits and loans. Consider the following situations:

- You hold \$10m worth of 91-day T-bills, which you will sell in one month because you will then need cash to pay your creditors. You fear a rise in interest rates over the next month, which will cause a fall in value of your T-bills. You can hedge your spot position in T-bills by shorting (selling) T-bill futures contracts.
- You will receive \$10m in six months' time, which you will then want to invest in a Eurodollar bank deposit for 90 days. You fear a fall in interest rates over the next six months, which means you will earn less interest on your deposits. You can hedge this risk by going long (i.e. buying) a Eurodollar futures contract.

- You hope to issue \$10m of 180-day commercial paper in three months' time (i.e. to borrow money). You fear that interest rates will rise over the next three months, so your borrowing costs will increase. You can hedge your borrowing costs by shorting (i.e. selling) T-bill futures contracts. This is an example of a cross-hedge, since there is no commercial bill futures contract so you have to use T-bill futures.

The key feature to remember when hedging using interest-rate futures is that, when interest rates (yield or discount rates) fall, the futures price will rise (and vice versa). However, we have not been precise about what interest rates we are referring to. To simplify matters, we assume rates (yields) for all maturities move together, that is, a parallel shift in the yield curve.

I.B.3.6.5 Hedge Ratios

We need to know exactly how many futures contracts to buy or sell, to minimise the hedging error. As we have seen, if the percentage change in the spot price (i.e. here the price of the spot T-bill) is greater than that of the futures price, then beta β_b will be greater than 1 and more futures contracts will be needed for the hedge to be successful, hence:

$$N_f = -\left(\frac{TVS_0}{FVF_0}\right)\beta_b$$

We could estimate β_b directly by regressing the percentage change in the underlying spot T-bill price, S , on the percentage change in the futures price, F , to give the *price sensitivity hedge ratio*. However, in fixed-income markets we are more used to thinking in terms of yield changes and linking these to price changes, using duration. We can in fact replace β_b in the above equation with an approximate expression incorporating duration, to give the *duration-based hedge ratio*. Using our duration formulae, with continuously compounded rates (see Chapter I.B.2):

$$\Delta S / S \approx -D_s \Delta y_s$$

$$\Delta F / F \approx -D_f \Delta y_f$$

where Δy_s is the change in the spot yield (on the underlying asset), Δy_f is the change in the forward rate, D_s is the duration of the physical asset and D_f is the duration of the futures contract. Hence:

$$N_f = \frac{TVS_0}{FVF_0} \left[\frac{D_s}{D_f} (\beta_y) \right] = \frac{TVS_0}{FVF_0} [\text{Rel.Vol.}] \quad (\text{I.B.3.5})$$

where the final term in square brackets refers to *relative volatility* (of the spot relative to the future) and an estimate of β_y is obtained from the regression $\Delta y_s = \alpha_0 + \beta_y \Delta y_f + \epsilon$. Equation (I.B.3.5) is

the *duration-based hedge ratio* (see Cuthbertson and Nitzsche (2001a) for further details). If spot and futures yields are assumed to move one-for-one (i.e. parallel shifts in the yield curve) then $\beta_y = 1$. Relative volatility is then simply the ratio of the durations and the equation simplifies to:

$$N_f = \frac{TVS_0}{FVF_0} \times \frac{D_s}{D_f}$$

For a wide variety of portfolios of debt instruments and holding periods, Toevs and Jacob (1986) looked at the performance of various hedge ratios. A basic one would be to use $N_f = (TVS_0 / FVF_0)$. Toevs and Jacobs found that the hedge ratio based on duration performed better than the ‘basic’ ratio. The above formula for the duration-based hedge ratio assumes that interest rates are continuously compounded. If we use discrete compounding, equation (I.B.3.5) becomes:

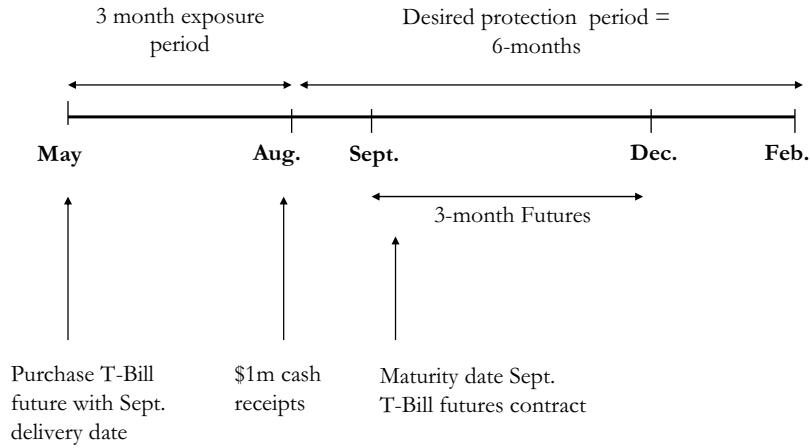
$$N_f = \frac{TVS_0}{FVF_0} \left[\frac{D_s (1 + y_f)}{D_f (1 + y_s)} (\beta_y) \right] = \frac{TVS_0}{FVF_0} [\text{Rel.Vol.*}]$$

The above equation, which uses discrete compound rates, and equation (I.B.3.5), which uses continuously compounded rates to calculate duration, generally give very similar values for N_f and both are approximations (since duration approximates the nonlinear relationship between the change in price and the change in yield). In real-world situations the above formula is usually used because in practice discrete compounding is the norm. However, note that as ‘modified duration’ is defined as $MD_i = D_i / (1 + y_i)$ for $i = s$ or f (see Section I.B.2.6) the above equation then ‘looks like’ (I.B.3.5). In what follows we treat both formulae for N_f as equivalent.

I.B.3.6.6 Hedging Using US T-bill Futures

Hedging with US T-bill futures is a little complex because of the difference between price *quotes* (as appear in the *Wall Street Journal*) and the *invoice price or face value of one contract*. Our example will be of a cross-hedge. Suppose it is May, and the treasurer of company X expects to receive \$1m in August and wishes to invest this in a six-month T-bill until February (see Figure I.B.3.1). The treasurer fears a fall in interest rates, which would imply that her investment beginning in August will earn less interest (or equivalently T-bills will cost more). A fall in rates implies a rise in prices so the treasurer buys (goes long in) T-bill futures so that the profit on the futures will compensate for the loss of interest on her spot market investment. The company hedges with a three-month T-bill future with a maturity date in September (i.e. closest to, but longer than, the hedge period May–August). This is a type of cross-hedge because the asset underlying the futures contract (i.e. three-month T-bill) does not match the underlying asset the treasurer wishes to invest in (i.e. six-month T-bill).

Figure I.B.3.1: Hedge using US T-bill futures



Example I.B.3.12

If the futures IMM index quote (in the *Wall Street Journal*) on the September T-bill futures contract is $Q_f = 89.2$, the futures price (per \$100 nominal) and the face value of one contract are given by:

$$F_0 = 100 - (100 - Q_f)/4 = 97.30$$

$$FVF_0 = \$1m \times F_0 / 100 = \$973,000$$

The price sensitivity hedge ratio (assuming a parallel shift in the yield curve, $\Delta y_s = \Delta y_f$) is:

$$N_F = \frac{TVS_0}{FVF_0} \frac{D_s}{D_f} = \frac{\$1m}{\$973,000} \left(\frac{0.5}{0.25} \right) = 2.05 \quad (2 \text{ contracts})$$

Between May and August the (six-month) spot yield falls (see Table I.B.3.3) implying that instead of earning a yield of $y_0 = 11\%$ the treasurer earns a yield of only $y_1 = 9.6\%$, since funds must be invested in August at the lower spot rate. However, lower spot rates imply that the implicit three-month forward rate also falls and hence futures prices rise (see Table I.B.3.3). The latter offsets some of the loss in the spot market.

Table I.B.3.3: Cross-hedge using US T-bill futures

<i>3 month US T-bill Futures: Sept Maturity</i>				
	Spot Market (May) (T-bill yields)	CME Index Quote Q_f	Futures Price, F (per \$100)	Face Value of \$1m Contract, FVF
May	y_0 (6m) = 11%	$Q_{f0} = 89.2$	97.30	\$973,000
August	y_1 (6m) = 9.6%	$Q_{f1} = 90.3$	97.58	\$975,750
Change	-1.4%	1.10 (110 ticks)	0.28	\$2,750 (per contract)
Durations are: $D_s = 0.5$, $D_f = 0.25$ Amount to be hedged = \$1m. No. of contracts held = 2				

Gain on the futures position

$$= (FVF_1 - FVF_0) N_f = z (F_1/100) - (F_0/100) N_f$$

$$= (\$975,750 - \$973,000) 2 = \$1m(0.97575 - 0.973) 2 = \$5,500$$

Gain on the futures position (using tick value of \$25 and $\Delta Q = 0.01$ is '1 tick')

$$= 110 \text{ ticks} \times \$25 \times 2 \text{ contracts} = \$5,500$$

Invest this profit of \$5,500 for six months (August–February) at $y_1 = 9.6\%$

$$= \$5,500 [1 + (0.096/2)] = \$5,764$$

Lost of interest in six-month spot market: unbedged ($y_0 = 11\%$, $y_1 = 9.6\%$)

$$= \$1m [0.11 - 0.096] (1/2) = \$7,000$$

Net loss: hedged position

$$= \$7,000 - \$5,764 = \$1,236$$

Had the company remained unhedged, the loss of interest would have been \$7,000 rather than the hedged loss of \$1,236. An alternative way of assessing the success of the hedge is to calculate the implied interest rate paid when hedged. The gain on the futures position of \$5,500 when invested over six months is \$5,764 hence (using simple interest)³:

$$\text{Effective ('simple') interest} = y_1 + 2 \times \$5764/\$1m = 0.096 + 0.0115 = 0.1075 = 10.75\%$$

³ To calculate the effective (simple) interest rate, expressed as an annual rate we have to add the gains on the futures position to the 6 months interest rate in August. The 6 months return on the futures position is \$5,764 (or \$5,764/\$1m over 6 months).

The 10.75% hedged return is substantially above the unhedged rate $y_2 = 9.6\%$ and is reasonably close to the implied (simple) yield on the September futures contract of 11.1% ($= (100/97.30 - 1)4$). The hedge is not perfect because the futures price and spot price are not perfectly correlated and the shift in the yield curve may not be parallel.

I.B.3.6.7 Summary

Short-term interest rate futures are extremely useful in locking in the yield to be paid to an investor between two time periods in the future. The optimal number of futures contracts for hedging purposes requires adjustments for:

- the size of the spot (cash) market position relative to the contract size of the futures;
- the duration of the spot market position to be hedged, relative to the duration of the asset underlying the futures contract; and
- the relative movement in the spot market interest rate and the rate underlying the futures contract; these adjustments give rise to the duration-based hedge ratio.

I.B.3.7 T-bond Futures

Some of the practical details of the use of T-bond futures are quite intricate. But, as with all futures contracts, T-bond futures can be used for speculation, arbitrage and hedging. A long T-bond futures position allows the holder to take delivery of a long-maturity T-bond at expiration. Hedging allows the investor to eliminate the price uncertainty of her bond portfolio. For example, suppose you hold $S = \$20\text{m}$ in T-bonds and fear a rise in long rates over the next six months. If you are long in T-bonds you should short T-bond futures. If long rates do subsequently rise the value of your bond portfolio, S , will fall. But as S falls so does the futures price, F . Hence, as you are short in the futures, you can close out the futures position and make a profit, which may offset most of the loss in the cash market.

If a speculator thinks long rates will fall in the future (i.e. bond prices rise) then she can purchase a T-bond future, since a rise in the cash market price implies a rise in the futures price. The investor gains leverage by purchasing the futures contract rather than purchasing the bonds outright in the cash market, because for the futures she has to pay only the initial margin. Transaction costs (e.g. bid–ask spread, clearing and brokerage fees) in the futures market might also be lower than those in the cash market. Naked speculative positions in futures are highly risky.

I.B.3.7.1 Contract Specifications

Futures contracts on a number of government bonds are traded on several exchanges. The most liquid contracts are in US T-bonds (CBOT) quoted in US dollars, on Euro Bonds of France

(traded on MATIF) and Germany (called ‘Bunds’ and traded on EUREX), both quoted in euros, and finally on UK gilts (quoted in sterling on LIFFE). The less liquid markets are in Spanish bonds (traded in euros on MATIF in Paris) and in Japanese government bonds (traded on LIFFE and in Tokyo).

I.B.3.7.1.1 UK Long Gilt Futures Contract

The details of the UK long gilt future (on LIFFE) are given in Table I.B.3.4. The gilt stated in the contract is a notional one with a 7% coupon (with a maturity between 8.75 and 13 years). Surprisingly, this notional 7% gilt does not actually exist! However, as we shall see, it provides a benchmark from which to calculate the price of possible bonds for delivery (which do exist). The seller of the futures contract can decide the exact delivery date (within the delivery month) and exactly which bond she will deliver (from a limited set, designated by the exchange). In fact, the seller will select a bond that is known as the *cheapest to deliver* (CTD). The CTD bond can be shown to depend on the *conversion factor* (CF) for each possible deliverable bond (these concepts are explained below).

The contract size is $\mathcal{z} = \text{£}100,000$ (nominal), futures price quotes are expressed per $\text{£}100$ nominal and the tick size is $\text{£}0.01$ per $\text{£}100$ nominal (e.g. one tick is a move from $\text{£}92$ to $\text{£}92.01$). Hence the tick value is $\text{£}10$ per contract ($= \text{£}0.01 \times \text{£}100,000/100$). If on 27th July the September futures had a closing (settlement) price of $F_0 = \text{£}114.19$ per $\text{£}100$ nominal, one futures contract on a nominal $\text{£}100,000$ would have a face value of $FV/F_0 = \mathcal{z} (F_0 / 100) = \text{£}114,190$.

Table I.B.3.4: UK Long Gilt Future (LIFFE)

Contract Size	$\text{£}100,000$ nominal, notional gilt with 7% coupon
Delivery Months	March, June, September, December
Quotation	Per $\text{£}100$ nominal
Tick Size (Value)	$\text{£}0.01$ ($\text{£}10$)
Last Trading	11 a.m., two business days prior to the last business day in the month
Delivery Day	Any business day in delivery month (seller’s choice)
Settlement	List of deliverable gilts published by the exchange with maturities between 8.75 and 13 years.
Margin Requirements	Initial margin $\text{£}2000$, Spread margin $\text{£}250$

I.B.3.7.1.2 US T-bond Futures Contract

The principles underlying this contract (on the CBOT) are very similar to those for the UK gilt futures contract, except price quotes are in thirty-seconds of 1% (see Table I.B.3.5). Expiration months are March, June, September and December. Note that the *notional asset* to be delivered in the contract is an 8% coupon bond. However, in practice, the short can choose from around 30 different eligible bonds to deliver (which have a coupons different from 8% and the CF adjusts the delivery price to reflect the type of bonds actually delivered). The T-bond delivered must have at least 15 years to maturity (or not be callable for at least 15 years). *The tick size is 0.03125% (e.g. for a contract size of \$100,000 nominal the tick value is \$31.25 per contract).*

Table I.B.3.5: US T-bond Future (CBOT)

Contract Size	\$100,000 nominal, notional US Treasury bond with 8% coupon
Delivery Months	March, June, September, December
Quotation	Per \$100 nominal
Tick Size (Value)	1/32 (\$31.25)
Last Trading Day	Seven working days prior to last business day in expiry month
Delivery Day	Any business day in delivery month (seller’s choice)
Settlement	Any US Treasury bond maturing at least 15 years from the contract month (or not callable for 15 years)
Margins	\$5000 initial; \$4000 maintenance
Trading Hours	8 a.m. – 2 p.m. Central Time
Daily Price Limit	96/30 pnts \$3000

Example I.B.3.13

On 7th July, suppose the US T-bond futures settlement price quote (CBOT) for September delivery was ‘98-14’ (= 98 + 14/32) which corresponds to a price of \$98.4375 per \$100 nominal. With a contract size of \$100,000, the face value of one contract is $FV/F_0 = z (F_0/100) = z$ (\$98.4375 / \$100) = \$98,437.50.

I.B.3.7.2 Conversion Factor and Cheapest to Deliver

Before discussing hedging strategies we need to be clear about the use of the conversion factor and the concept of the cheapest-to-delivery bond. We discuss these with respect to the US T-bond future, but similar principles apply to the UK gilt future. There are a wide variety of bonds with over 15 years to maturity that can be delivered by the investor who is short in the futures

market. These will have different maturities and coupon payments, and hence different values for CF and the CTD.

The CF is best understood using a specific example. To simplify matters we assume the short does not yet hold a bond for delivery, so we will calculate the CF and CTD bond at the maturity date of the contract (T). Consider a 10% coupon, 20-year US T-bond paying semi-annual coupons of \$5 with $n = 40$ (six-month) periods to maturity. If the yield to maturity (YTM) on this bond were 8% then its ‘fair’ or ‘theoretical’ price would be \$119.794 (per \$100 nominal) (See Chapter I.B.2). This gives a conversion factor $CF_T = 1.19794$. In essence this deliverable bond is worth 1.19794 times as much as the notional 8% coupon bond (trading at par and hence with its YTM of 8%). The conversion factor adjusts the price of the actual bond delivered, relative to the notional 8% coupon bond in the futures contract and it is easy to see that *if the coupon on the bond actually delivered exceeds 8% then $CF > 1$* , whereas *if the coupon on the bond actually delivered is less than 8% then $CF < 1$* .

The conversion factor will differ for bonds with different coupon payments and different maturities. It will also change through time on any specific bond simply because the maturity date gets closer (although it will always exceed 15 years, otherwise it will cease to be an eligible bond).

Example I.B.3.14

Data: *Possible bond for delivery is a 10% coupon T-bond with remaining maturity of 20 years (semi-annual coupons)*

Question: *What is the conversion factor of the bond?*

Answer: The **theoretical price** of the bond if its YTM is assumed to be 8% (i.e. the same as the notional bond in the contract) is:

$$P = \frac{\$5}{(1.04)} + \frac{\$5}{(1.04)^2} + \dots + \frac{\$5}{(1.04)^{40}} + \frac{\$100}{(1.04)^{40}}$$

Using the annuity formula for the discounted present value of the coupons

$$P = 19.7928 \times \$5 + 0.2083 \times \$100 = \$119.79$$

The theoretical price of the actual bond to be delivered is $P = 119.79$ and hence the conversion factor $CF = 1.1979$ (per \$100 nominal).

Suppose it is now T (i.e. 11th September) and there are three bonds designated by CBOT for actual delivery in September whose CFs are 1.044, 1.033 and 1.065. In practice the calculation of the CTD bond is quite complex but we show below that a rough idea of the CTD bond can be obtained by choosing that bond with the smallest raw basis:

$$\text{Raw Basis} = B_T - F_T CF_T$$

where B_T is the spot ('clean') price of an eligible bond for delivery, F_T is the settlement futures price and CF_T is the conversion factor of a deliverable bond.

Table I.B.3.6 shows the calculation of the raw basis for three deliverable bonds. The final column indicates that bond C is the CTD. (For a more detailed account of the calculation of the CTD bond see Hull (2000).) Although it is often the case that the bond actually delivered by the short is the CTD, this is not always so (see Gay and Manaster 1986), since the seller may wish to preserve the duration of her own portfolio or provide a bond, other than the CTD, in order to minimise her tax bill.

Table I.B.3.6: The CTD bond

Deliverable Bonds (maturity)	Spot Price (S)	Conversion Factor (CF)	Raw Basis ($= S - F \times CF$)
1. Bond A (2019)	112-4 (112.125)	1.044	0.156
2. Bond B (2025)	112-8 (112.25)	1.033	1.461
3. Bond C (2027)	114-8 (114.25)	1.065	0.029

$$F \text{ (September Futures)} = 107-8 \text{ (107.25)}$$

I.B.3.7.3 Hedging Using T-bond Futures

The principles involved in hedging a T-bond portfolio are much the same as those when hedging a portfolio of T-bills, discussed earlier. If Ms A is long in bonds and fears a fall in bond prices, then to hedge the position she will go short in bond futures. Conversely, if she wishes to purchase bonds *in the future* and is worried that bond prices will rise (i.e. yields will fall), she should buy bond futures. The optimal number of futures contracts, N_f is given by the *duration-based hedge ratio* (see earlier discussion) with the added complication of the conversion factor for the CTD bond:

$$N_f = -\frac{TVS_0}{FVF_0} \left[\frac{D_s}{D_f} (\beta_y) \right] CF_{CTD} = -\frac{TVS_0}{FVF_0} [\text{Rel.Vol.}] CF_{CTD} \quad (\text{I.B.3.6})$$

where

TVS_0 is the total market value of portfolio of bonds to be hedged

$FVF_0 = (\$F/100)$ is the face value of one bond futures contract

S is the invoice spot/cash price of bond(s) to be hedged

F is the invoice price of the CTD bond in the futures contract

CF_{CTD} is the conversion factor of the CTD for the futures contract

D_s is the duration of spot bond to be hedged at the expiration of the hedge

D_f is the duration (at expiration of the futures) of the actual bond to be delivered in the contract

y_s is the YTM of the underlying bond to be hedged

y_f is the YTM of the bond in the futures contract

Note: If y_s and y_f are the *quoted* yield to maturity (e.g. when using discrete semi-annual compounding) then we use *modified* durations for D_s and D_f .

I.B.3.7.4 Hedging a Single Bond

Consider a US pension fund manager on 1st May who wishes to hedge her nominal bondholding of $NB_0 = \$1m$, held in a 10%, 2005 Treasury bond with current price $S_0 = 101-00$ (\$101 per \$100 nominal). The market value of the spot position is $TVS_0 = \$1,010,000$ ($= NB_0(S_0/100) = \$1m \times 101/100$). The fund manager fears a rise in interest rates over the next three months (from 1st May to 1st August), so the value of her bond portfolio may fall and she may not be able to meet scheduled pension payments. The US T-bond futures contract details are given in Example I.B.3.15 and the CF of the CTD bond is 1.12. The fund manager purchases the nearby September contract on 1st May. The optimal number of futures contracts N_f is given by equation (I.B.3.6) and assuming a parallel shift in the yield curve ($\Delta y_s = \Delta y_f$):

$$N_F = - \frac{TVS_0}{FVF_0} \left(\frac{D_s}{D_f} \right) CF_{CTD} \quad (\text{I.B.3.7})$$

$$= - \frac{\$1,010,000}{\$110,500} \times \frac{6.9}{7.2} \times 1.12 = -9.04 \times 0.958 \times 1.12 = -9.7$$

(i.e. short 10 contracts).

Sometimes the face value of the spot position is (erroneously) taken to be $B_0 = \$1m$ rather than $TVS_0 = B_0(S_0/100) = \$1,010,000$, but if the bond is selling near par this makes little difference. Suppose the fund manager's worst fears are realised and interest rates rise between 1st May and 1st August. As the spot bond price falls from $S_0 = 101-00$ to $S_1 = 98-16$, she makes a capital loss in the cash market of \$25,000, but the futures price moves from $F_0 = 110$ to $F_1 = 108-16$. The gain on the short futures position is \$20,000 ($= 2 \times 32 \text{ ticks} \times \$31.25 \times 10 \text{ contracts}$). The result of the hedge shows a small loss of \$5000 on a portfolio of a notional \$1m (i.e. about 0.5%). The unhedged portfolio would have lost \$25,000 (i.e. about 2.5%). If the fund manager does not wish to assume $\Delta y_s = \Delta y_f$ then a regression of Δy_s on Δy_f using past data provides an estimate of β_y .

Example I.B.3.15: Hedging a single bond

<u>Spot Position: 1st May</u>	<u>Futures: 1st May (September delivery)</u>
10%, 2005 Treasury bond (YTM = 10.12%)	CF of CTD bond = 1.12
Nominal bond holding $NB_0 = \$1m$	Size of one contract $z = \$100,000$
Current Price $S_0 = \$101$ (\$101 per \$100)	Price of futures $F_0 = 110-16$ (\$110.50)
Market Value, Spot = TVS_0	Face Value $FV/F_0 = zF_0 = \$110,500$
= $NB_0(S_0/100) = \$1,010,00$	Duration $D_f = 7.2$
Duration $D_s = 6.9$	Tick value equals \$31.25
Number of futures contracts	
Outcome of the Hedge: 1st August	
<u>Spot Market (On 1st August)</u>	<u>September Futures (On 1st August)</u>
$S_1 = 98-16$ (\$98.50)	$F_0 = 110-16$ (\$110.50)
Value of spot position $TVS_1 = (S_1/100)B$	$F_1 = 108-16$ (\$108.50 per \$100)
= $(98.50/100)\$1m$	
= \$985,000	
Loss on spot position = $((S_0 - S_1)/100)B$	Gain on short futures
= $(101 - 98.5)/100)\$1m = \mathbf{\$25,000}$	= $N_f z (F_0 - F_1)/100$
	= $10 (\$100,000) (2/100) = \mathbf{\$20,000}$, or
	= $(2 \times 32) \text{ ticks} \times \$31.25 \times 10 = \mathbf{\$20,000}$
<u>Net Loss</u>	
Hedged = $\$20,000 - \$25,000 = \$5,000$	
Unhedged = $\$25,000$	

I.B.3.7.5 Hedging a Portfolio of Bonds

For illustrative purposes assume a fund manager has two bonds in her portfolio, with \$10m in bond A and \$20m in bond B, so that (if both bonds are priced at par) $TVS_0 = \$30m$. Using equation (I.B.3.7), the only change required to calculate the number of futures contracts to short is to replace D_s by the weighted duration of the cash market bond position. The ‘new’ D_s is:

$$D_s = [w_A S_A D_A + w_B S_B D_B]$$

where $w_A = 0.33$ ($= \$10m / \$30m$) and $w_B = 0.67$ ($= \$20m / \$30m$), S_i is the market price of bond i (per \$100 nominal), D_i is the duration of bond i , for $i = A$ or B .

The hedged position will not provide a perfect hedge. However, numerous studies (e.g. Toevs and Jacob 1986, Gay et al. 1983, Chance et al. 1986) find that hedging using T-bond futures can reduce risk by about 50–70%, with the duration-based hedge ratio performing particularly well. Hedging cash positions in *corporate* bonds using *T-bond* futures and using the duration-based hedge ratio is also found to be highly effective, even though this is a cross-hedge.

I.B.3.7.6 Summary

- The contract details for T-bond futures contracts are necessarily complex because of the use of the conversion factor and the cheapest-to-deliver bond.
- T-bond futures can be used to hedge an underlying government bond portfolio by using the duration-based hedge ratio to determine the number of futures contracts to short. Conversely, hedging the purchase of bonds at some time in the future requires that you enter today into a long T-bond futures position.
- T-bond futures contracts can also be successful in hedging an underlying corporate bond portfolio even though this is a cross-hedge.

I.B.3.8 Stack and Strip Hedges

We have seen how to determine the optimal number of futures contracts in order to hedge changes in interest rates. Exactly which contracts we use will be determined by the liquidity of the different contracts and the desired period(s) over which we wish to hedge our exposure. For example, suppose it is 10th December and a corporate has a bank loan for \$10m principal, with interest-rate reset dates every three months on 10th March, 10th June and 10th September, and the loan is repaid the following 10th December. The corporate is worried that interest rates will rise over the coming year. To hedge the position the corporate needs to short $N_f = \$10\text{m}/\$1\text{m} = 10$ Eurodollar futures contracts for each of the three reset dates. What maturity contracts should the corporate use?

The corporate could undertake a *strip hedge*, whereby on 10th December it shorts 10 March contracts, 10 June contracts and 10 September contracts using 90-day Eurodollar contracts. The corporate closes out each of these contracts at the time the interest-rate resets occur. If interest rates rise, any profit from closing out the maturing futures contracts can be used to offset the higher interest payments payable on the bank loan. Hence, on 10th December, the corporate effectively ‘locks in’ the futures interest rates that apply between March and June (f_{12}), June and September (f_{23}), and September and December (f_{34}).

An alternative is for the corporate on 10th December to undertake a *stack hedge*. Here the corporate sells 30 of the three-month Eurodollar futures that mature at the end of September,

after the last reset date for the bank loan. If interest rates increase between 10th December and 10th March, the short futures position will earn a profit, which effectively reduces the higher interest the corporate will pay on its bank loan over the next three months. Immediately after 10th March the corporate therefore buys back 10 September futures contracts, leaving it with 20 short futures contracts to hedge the next two interest-rate resets. On 10th June the corporate will buy back a further 10 of the September contracts, leaving it short the remaining 10 contracts, to hedge the final interest-rate reset on 10th September. The corporate is locking in the forward rate on the September contract (f_{34}) on each of the reset dates.

Swap dealers frequently hedge their interest rate exposure using Eurodollar (Euribor, etc.) contracts; these are usually fairly liquid markets out to at least five years, and execution of the above stack or strip hedge is possible.

What are the relative merits of a stack versus a strip hedge? The stack hedge uses a *single* futures contract (with September futures rate f_{34}) to hedge *all* the future interest-rate resets. If movements in the yield curve are parallel then the (three-month) futures rate underlying the contract f_{34} will move the same as f_{12} and f_{23} , and the stack hedge will be effective. However, if the forward rate f_{34} moves differently from f_{23} and f_{34} (i.e. a twist in the yield curve), then the stack hedge may not provide an efficient hedge for future interest-rate resets. This is the danger in a stack hedge.

Another example of stack and strip hedges is provided by the need to hedge a possible interest-rate rise between, say, 1st January and 10th March, when you will take out a \$10m bank loan with a *one-year* floating rate based on 90-day LIBOR. The underlying in the Eurodollar futures is a 90-day interest rate, but the bank loan has a one-year floating rate. The duration-based hedge ratio indicates that you should short $N_f = (\$10m/\$1m) \times 12/3 = 40$ contracts for this one interest-rate reset. You could use 40 March contracts and this would be a stack hedge. Alternatively, you could use 10 each of the March, June, September and December contracts – this would be a strip. In both cases you close out all contracts on 10th March, when you take out the bank loan. The strip locks in a forward rate that is the average of the forward rates in the four separate contracts, whereas the stack hedge locks in the forward rate on the March contract only. Again, if movements in the yield curve are parallel then there is no difference in the two types of hedge. But if the ‘far’ futures prices move more than the ‘nearby’ futures rates (i.e. a twist in the yield curve), the strip hedge will be more effective in locking in the one-year rate than the stack hedge.

The advantage of the stack hedge is that it uses the nearby contract, which may be more liquid than the longer-maturity contracts, but is subject to more basis risk than the strip hedge. If the

hedge period is actually longer than the maturity of the *longest* futures contract then there is an additional problem. Suppose you have agreed to supply heating oil at a fixed price in 15 years' time but the liquidity in heating oil futures contracts goes out only five years. Then you can hedge using only the five-year futures contract and after that period you have to enter into another five-year hedge at whatever futures price then prevails. In this simplified example you are *rolling over the hedge* every five years and this introduces basis risk, when hedging price changes for longer than a five-year horizon. In 1994 this was the problem faced by Metallgesellschaft, a German company.

I.B.3.9 Concluding Remarks

In this chapter we have shown how a wide variety of spot (cash) market assets such as a diversified portfolio of stocks, currencies, commodities and interest-sensitive assets such as T-bills and T-bonds can be hedged using the appropriate futures contract. The actions of arbitrageurs keep changes in the futures price closely aligned to changes in the spot price and it is this that allows a successful hedging strategy. When hedging, a near-perfect negative correlation between the spot price S and the futures price F is produced by taking an opposite position in the two assets (i.e. a long-short position). The hedge is never perfect because of basis risk, but the latter is generally relatively small. Hence, hedging can be viewed either as losses in the spot market being offset by gains in the futures market or as the final spot price plus any gains/losses in the futures position 'locking in' a price close to the originally quoted futures price F_0 . Finally, speculation with futures is straightforward. The futures price and the price of the underlying asset are highly positively correlated, hence a speculator will go long in (i.e. buy) a futures contract if she thinks the spot price of the underlying asset will rise (and vice versa) and then close out the position at a later date. If she has made the forecast correctly then she will receive net cash receipts from the futures clearing house (see Chapter I.C.6 for institutional details of margin accounts).

References

Baillie, R T, and Myers, R J (1991) 'Bivariate GARCH estimation of the optimal commodity futures hedge', *Journal of Applied Econometrics*, 6, pp. 109–24.

Chance, D, Marr, M W, and Thompson, G R (1986) 'Hedging shelf registrations', *Journal of Futures Markets*, Vol. 6, pp. 11-27.

Cuthbertson, K, and Nitzsche, D (2004) *Quantitative Financial Economics: Stocks, Bonds and Foreign Exchange* (Chichester: Wiley).

Cuthbertson, K, and Nitzsche, D (2001a) *Investments: Spot and Derivatives Markets* (Chichester: Wiley).

Cuthbertson, K, and Nitzsche, D (2001b) *Financial Engineering: Derivatives and Risk Management* (Chichester: Wiley).

Gay, G D, Kolb, R W, and Chiang, R (1983) 'Interest rate hedging: An empirical test of alternative strategies', *Journal of Financial Research*, Vol. 6, pp. 327-343.

Gay, G D and Manaster, S (1986) 'Implicit delivery options and optimal delivery strategies for financial futures contracts', *Journal of Financial Economics*, Vol. 15, pp. 41-73.

Hull, J C (2000) *Options, Futures and Other Derivatives* (London: Prentice Hall)

Toevs, A, and Jacob, D (1986) 'Futures and alternative hedge methodologies', *Journal of Portfolio Management*, Spring, pp. 60–70.

I.B.4 Swaps

Salih N. Neftci*

Interest rate swap (IRS) markets rank among the worlds largest in nominal amount and are among the most liquid. The size of the world’s various swap markets is shown in Figure I.B.4.1. We see that swap market activity dwarfed popular markets such as stock and bond markets in mid-2002. Why is this so?

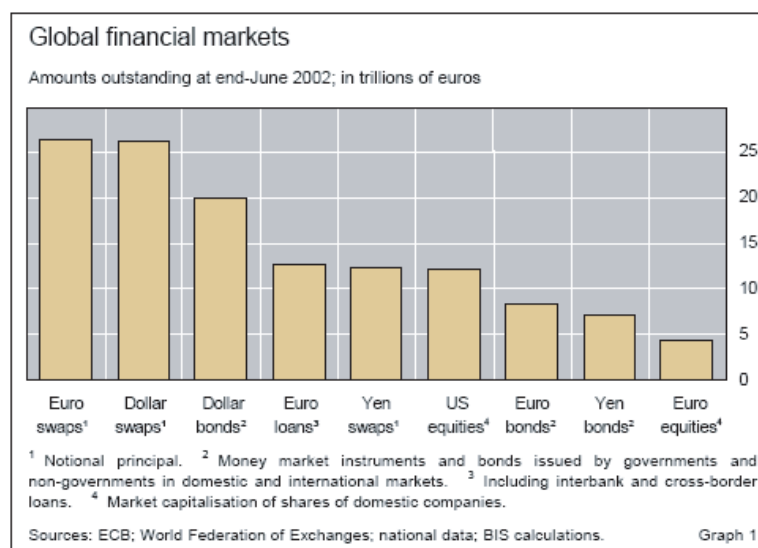


Figure I.B.4.1: Swap Market Activity.

There are many uses for swaps. Borrowers arbitrage credit spreads and borrow currencies that yield the lowest all-in-cost. This, in general, implies borrowing in a currency other than the one the issuer needs. The proceeds, therefore, need to be swapped into the needed currency and may also be swapped from fixed rate to floating rate. Hence, new bond issues are one source of liquidity for swap markets.

Balance sheet management of interest exposure is another reason for the high liquidity of swaps. The asset and liability interest rate exposure of financial institutions can be adjusted using interest rate swaps and swaptions. If funding is obtained in floating rates and on-lent at a fixed rate, then an interest rate swap can be entered into and the exposures can be efficiently managed.

Swap market liquidity is further supported by the needs of mortgage based hedging activity. It appears that a significant portion of plain vanilla swap trading is due to the requirements of the mortgage

*Salih N.Neftci is a Professor of financial economics at the Graduate School of City University, New York, and Head of the FAME Certificate program in Geneva.

industry. In the US, for example, most mortgages carry a fixed rate of interest. They have prepayment clauses and this introduces convexity in mortgage portfolios. This convexity (reference IB2) can be hedged using swaptions,(reference IB7) which creates liquidity in the market for both swaps and swaptions.

It turns out that swaptions are option positions that need to be dynamically hedged. This hedging can be done with forward swaps as the underlying which leads to further swap trading. Mortgage markets are huge and this activity can sometimes dominate the swap and swaption markets.

Swaps exist in many varieties, only some of which can be explained in detail in this chapter. Our focus here is on two of the most traded swap varieties - interest rate swaps and currency swaps. For both instruments we explain their features, their uses and their pricing. We also consider the structure and uses of equity swaps, commodity swaps, volatility swaps and basis swaps. There are several ‘exotic’ swap varieties that we briefly discuss. Credit default swaps are examined in I.B.6 along with other credit derivatives. While FRAs can be analyzed using a swaps framework, they are examined in I.B.3 along with futures and forwards.

I.B.4.1 What is a swap?

Imagine *any* two cash flows with different characteristics. One can, then, devise a contract to exchange these cash flows. This contract will be called a *swap*. In general, a swap contracts has the following characteristics.

1. A swap is a pure exchange and hence should, in principle, not require any additional net cash payments at initiation. In other words the *initial* value of the swap contract that secures this exchange should be *zero*.
2. The contract needs to specify a *swap rate*, or *spread* to make the two swap parties willing to exchange the cash flows.

A generic exchange is shown in Figure I.B.4.2 (a). In this Figure the first series of cash flows starts at time t_1 and continues periodically at t_2, t_3, \dots, t_k . There are k cash flows of differing sizes denoted by:

$$\{C(s_{t_o}, x_{t_1}), C(s_{t_o}, x_{t_2}), \dots, C(s_{t_o}, x_{t_k})\} \quad (\text{I.B.4.1})$$

These cash flows depend on a vector of risk factors denoted by x_{t_i} . These latter could be various market and credit risks. But, the cash flows depend also on a variable s_{t_o} decided at time t_o . The s_{t_o} is either a swap *spread* or an appropriate *swap rate*. By selecting the ‘correct’ value of s_{t_o} the initial value of the swap will be made zero. Figure I.B.4.2 (b) represents another strip of cash flows:

$$\{B(y_{t_1}), B(y_{t_2}), \dots, B(y_{t_k})\} \quad (\text{I.B.4.2})$$

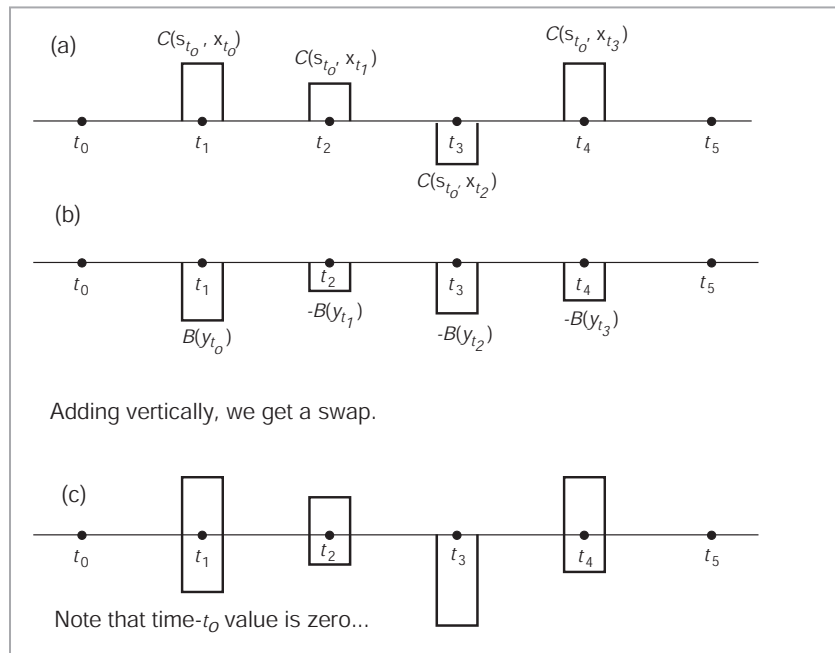


Figure I.B.4.2: A Generic Swap.

which depends potentially on some other risk factors denoted by y_{t_i} .

The swap consists of exchanging the $\{C(s_{t_o}, x_{t_i})\}$ against $\{B(y_{t_i})\}$, at *settlement dates* $\{t_i\}$. The ‘price’ s_{t_o} is selected at time t_o so that the swap parties are willing to go through with this exchange with no initial cash payment. This is shown in Figure I.B.4.2 (c). One party will *pay* the $C(\cdot)$ ’s and *receive* the $B(\cdot)$ ’s. The counterparty will be on the ‘other side’ of the deal and will do the reverse. Clearly, if the cash flows are in the *same* currency, then there is no need to make two different payments at each period t_i . One counterparty can simply pay the other the *net* amount.

Example I.B.4.1 *Suppose you sign a contract that entitles you to a 7% return in dollars, in return for a 6% return in Euros. The exchanges will be made every 3-months based on nominal amounts determined at a pre-determined exchange rate e_{t_o} . At initiation time t_o , the net value of the commitment should be zero, given the correct swap spread. This means that the market value of the receipts and payments will be identical at time t_o , when evaluated at e_{t_o} . Yet, once the contract is signed, USD interest rates may fall relative to European rates. This would make the receipt of 7% USD funds relatively more valuable than the payments of 6% in Euro, assuming no change in the exchange rate.*

As a result, from the point of view of the USD-receiving party, the value of the swap will move from zero to positive, while for the other counterparty the swap will have a negative value. Hence one counterparty benefits, and the one who made the opposite commitment loses exactly the same amount.

I.B.4.2 Types of Swaps

Swaps are a very broad instrument category. Practically any sequence of cash flows can be used to generate a swap. Most swaps are interest rate related given the Libor and yield curve exposures on corporate and financial balance sheets. But, swaps form a very broad category of instruments and to emphasize this point we start the discussion with non-interest rate based swaps.

I.B.4.2.1 Equity Swaps

One interesting swap category is the exchange of returns from an equity or equity index, against the return of another asset, often against Libor based cash flows. Such instruments are called *equity swaps*.

An equity swap will be initiated at time t_o . An equity index I_{t_i} , and a money market rate, say Libor, L_{t_i} are selected. At times $\{t_1, t_2, \dots, t_n\}$ the swap parties will exchange cash flows based on the percentage change in I_{t_i} , written as:

$$N_{t_{i-1}} \left(\frac{I_{t_i} - I_{t_{i-1}}}{I_{t_{i-1}}} \right)$$

against Libor based cash flows, $N_{t_{i-1}} L_{t_{i-1}} \delta$, where the δ will represent the day's adjustment since the i may run across quarters, for example. The N_{t_i} is the notional amount, which is not exchanged. The notional amount can be changed at times t_1, \dots, t_{n-1} allowing the parties to adjust their position in the particular equity index periodically.

Example I.B.4.2 Suppose the notional amount is 1 million USD and we have the following data:

$$I_{t_o} = 800$$

$$I_{t_1} = 850 \quad L_{t_o} = 5\%p.a. \quad spread = .20\%p.a.$$

Then the time t_1 equity-linked cash flow is:

$$1m \left(\frac{I_{t_1} - I_{t_o}}{I_{t_o}} \right) = 1,000,000(0.0625) = 62,500$$

the Libor linked cash flow will be:

$$1m(L_{t_o} - s_{t_o}) \frac{90}{360} = 1,000,000(.05 - .002) \frac{1}{4} = 12,000$$

The remaining unknown cash flows will become known as time passes, the dividends are paid, and prices move.

Putting two equity swaps together one can swap returns between *two* equity indices. For example one can buy one equity swap for Nasdaq, and sell one for S&P500 indices. The Libor-related cash flows would then cancel and the two parties would be exchanging two returns directly. A spread would make sure that the exchange of cash flows is accepted by both parties without any upfront payment.

The pricing of equity swaps depends on the characteristics of the underlying equity index. Assuming that risk-free rate is constant during the relevant period, we can use risk-neutral probabilities \tilde{P} to obtain the relation:

$$E_{t_o}^{\tilde{P}} \left[\frac{I_{t_i} - I_{t_{i-1}}}{I_{t_{i-1}}} \right] = (r - d)(t_i - t_{i-1})$$

where the d is the constant and known rate of dividend payments. The $(t_i - t_{i-1})$ is the δ parameter mentioned above. If the index is already purged of dividend payments then the d will be zero. The pricing of the equity swap will be based on this relationship. Supply-demand, differences between Libor and r , and the way dividends are taken into consideration will determine the spread asked.

I.B.4.2.2 Commodity Swaps

The overall structure of commodity swaps is similar to equity swaps. There are two major types of commodity swaps. Swap parties can either, (1) exchange fixed to floating payments based on an commodity price index, or (2) exchange payments when one payment is based on an index and the other on a money market rate, which is often the Libor rate.

Example I.B.4.3 *Consider for example, a refinery. The refinery buys crude oil and sells refined oil products. This refinery may find it useful to lock in a fixed price for crude oil.*

Hence this refiner may want to receive a floating price of oil and pay a fixed price. When coupled with spot purchases of oil, this swap will eliminate the floating oil price risk for the refinery. The floating oil price that is received would then be paid to buy the spot oil. The risk associated with oil price movements would disappear and the refinery would achieve a fixed oil price.

Such *commodity swaps* can be arranged for all sorts of commodities, metals, precious metals, and energy prices. Pricing of commodity swaps can again be based upon the risk-neutral measure. Under the assumption that the underlying commodity is actively traded in the markets, the price of the commodity will increase at the risk free rate if there are no convenience yield, no storage cost etc...

I.B.4.2.3 Interest Rate Swaps

This is the most widely used swap category. It involves exchanging cash flows generated by different interest rates. The most common case is where a *fixed* swap rate is paid against the receipt of a floating Libor rate, in the *same* currency. A *plain vanilla interest rate swap* (IRS) initiated at time t_o is a commitment to exchange interest payments associated with a notional amount N , settled at clearly identified settlement dates, $\{t_1, t_2, \dots, t_n\}$. The *buyer* of the swap will make fixed payments of size $s_{t_o} N \delta$ each, and receive floating payments of size $L_{t_i} N \delta$. The Libor rate L_{t_i} will be determined at *set dates* $\{t_o, t_1, \dots, t_{n-1}\}$. The maturity of the swap will be m years.¹

¹Here $m = n\delta$.

Example I.B.4.4 An IRS has a notional amount N of 1million USD. A 7% fixed rate for 2 years is paid in semi-annual (s.a.) payments against a cash flow generated by 6-month Libor. This is shown in Figure I.B.4.3 (a). There are two cash flows. One involves four payments of 35,000USD each. This is known at t_0 , and paid at the end of each 6-month period.

The second is shown in Figure I.B.4.3 (b). It will be determined by the value of 6-month USD Libor to be observed at set dates. Since there are 2 years, 4 separate Libor rates will be observed during this period. The first Libor L_{t_0} is known at the initial point t_0 . The remaining Libor rates $L_{t_1}, L_{t_2}, L_{t_3}$ will be observed gradually as time passes, but are unknown initially.

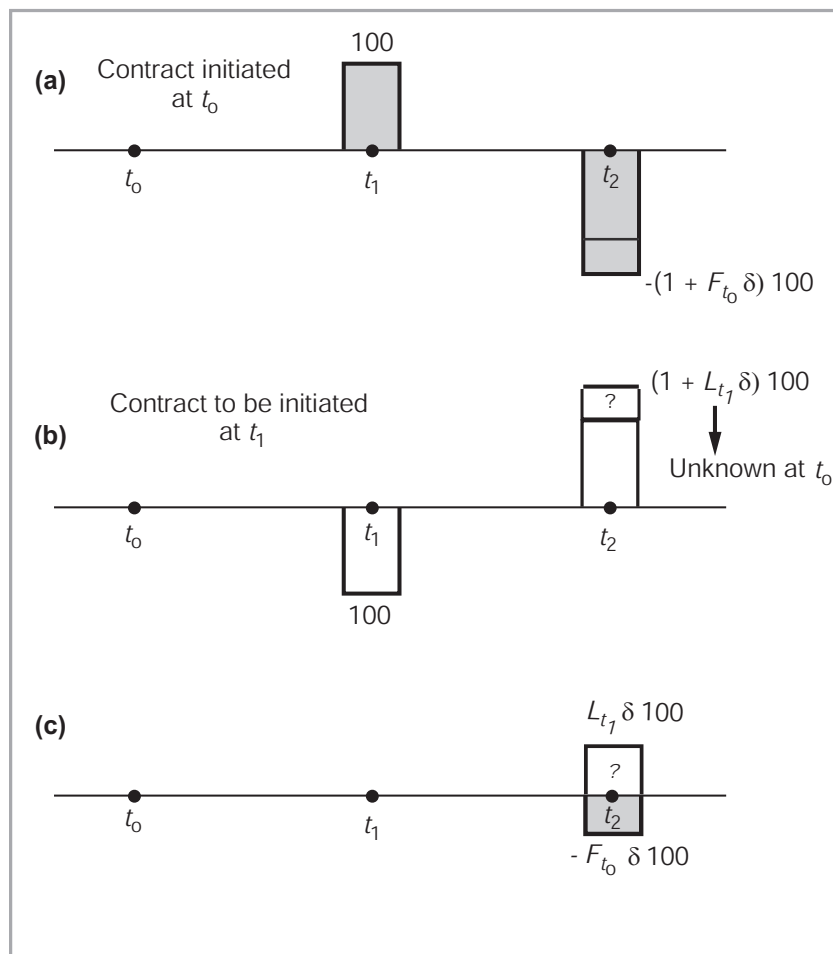


Figure I.B.4.3: An interest Swap.

The pricing of interest rate swaps is examined in I.B.4.5.

I.B.4.2.4 Currency Swaps

Currency swaps are similar to interest rate swaps but there are some major differences. First, the exchanged cash flows are in *different currencies*. This means that two different yield curves are involved

in swap pricing instead of just the one, as in the case of an interest rate swap. Second, in the large majority of cases a floating rate is exchanged against another floating rate. A third difference is that the principals are exchanged at initiation and re-exchanged at maturity. In the case of interest rate swaps this question does not arise since it would amount to exchanging N against N , the notional amounts being in the same currency.

Since currency swaps are otherwise very similar to interest rate swaps, the engineering of currency swaps will be done in almost the same way. Formally, a currency swap will have the following components. There will be two currencies, say USD(\$) and Euro(€). The swap will be initiated at time t_o and will involve (1) an exchange of a principal amount $N^{\$}$ against the principal $M^{\text{€}}$, (2) a series of floating interest payments associated with the principals $N^{\$}$, and $M^{\text{€}}$ respectively, at settlement dates, $\{t_1, t_2, \dots, t_n\}$. One counterparty will pay the floating payments $L_{t_i}^{\$} N^{\$} \delta$ each, and receive floating payments of size $L_{t_i}^{\text{€}} M^{\text{€}} \delta$. The two Libor rates $L_{t_i}^{\$}, L_{t_i}^{\text{€}}$ will be determined at set dates $\{t_o, t_1, \dots, t_{n-1}\}$. The maturity of the swap will be m years. A small spread of s_{t_o} will be added to one of the interest rates in order to make both parties exchange the cash flows. The market maker will quote bid/ask rates for this spread.

Example I.B.4.5 Figure I.B.4.4 shows a currency swap. The USD notional amount is 1million. The current EUR/USD exchange rate is at .95. The agreed spread is 6 basis points(bp) over USD Libor as it is customary in this market. The initial 3-month Libor rates are:

$$L_{t_i}^{\$} = 3\%$$

$$L_{t_i}^{\text{€}} = 3.5\%$$

This means that at the first settlement date

$$(1,000,000)(.03 + .0006\frac{1}{4})\frac{1}{4} = \$8250$$

will be exchanged against,

$$(1,000,000)(.95)(.035)\frac{1}{4} = \text{€} 8312.$$

all other interest payments would be unknown. Note that the Euro principal amount is related to the USD principal amount according to:

$$N^{\$} e_{t_o} = M^{\text{€}}.$$

Also, note that we added the swap spread to the USD Libor.

Pricing currency swaps will follow the same principles as in the case of interest rate swaps.

A currency swap involves two streams of cash flows in two different currencies. One can then reverse engineer a currency swap as shown in Figure I.B.4.4. According to this Figure a currency swap is

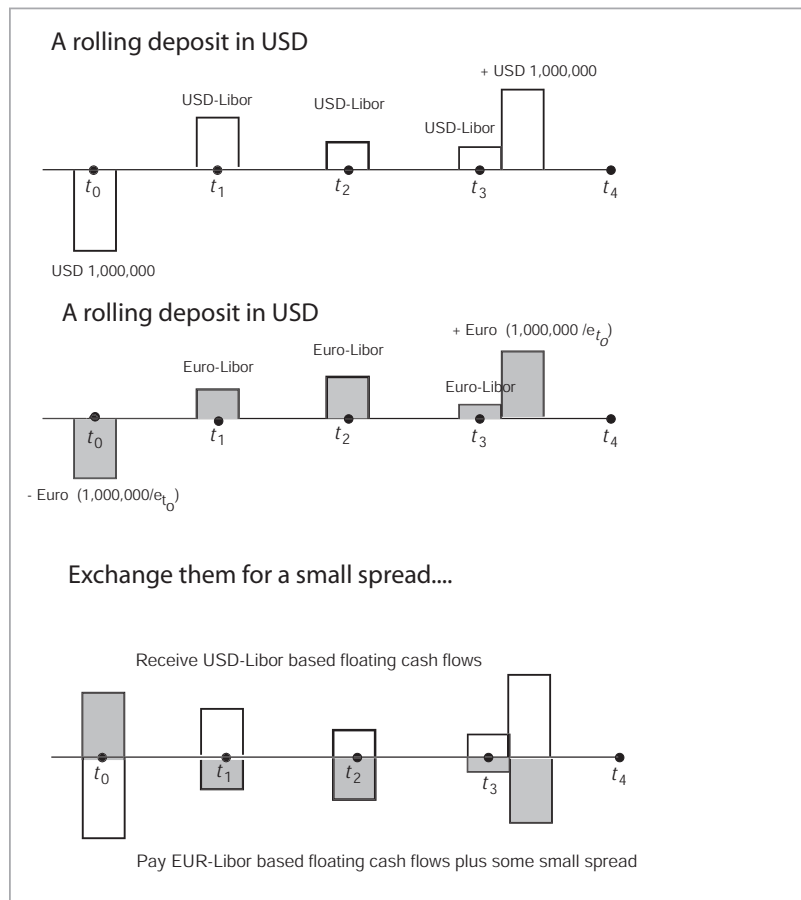


Figure I.B.4.4: A currency Swap.

equivalent to a swap of two floating rate deposits rolled over at regular intervals of length δ . Similarly, one can visualize the currency swap as a swap of two FRN's in different currencies that reset at the same times t_i .

Once this engineering is done we know that at each reset date t_i , the present value of the future cash flows will always equal the underlying notional amount. Hence for a currency swap that ends at time t_n we have at each $t_i < t_n$,

$$N^{\$} = \frac{L_{t_i}^{\$} \delta N^{\$}}{(1 + L_{t_i}^{\$} \delta)} + \dots + \frac{(1 + L_{t_{n-1}}^{\$} \delta) N^{\$}}{(1 + L_{t_i}^{\$} \delta) \cdots (1 + L_{t_{n-1}}^{\$} \delta)}$$

This will be the case since we have for the last term on the right hand side

$$\frac{(1 + L_{t_i}^{\$} \delta) N^{\$}}{(1 + L_{t_{n-1}}^{\$} \delta) \cdots (1 + L_{t_{n-1}}^{\$} \delta)} = \frac{N^{\$}}{(1 + L_{t_i}^{\$} \delta) \cdots (1 + L_{t_{n-2}}^{\$} \delta)}$$

This way, all terms involving the unknown Libor rates will cancel recursively, and the present value of the future cash flows will always equal $N^{\$}$ for the US dollar leg, and the $N^{\text{€}}$ for the Euro Libor leg.

The implication of this is the following. For major currencies, exchanging two FRN's at times t_i is equivalent to exchanging the underlying notionals. At the time of initiation of the currency swap this exchange will be done at the prevailing spot exchange rate e_{t_0} . Thus, the currency swap spread will in theory be close to zero.

However, in practice this spread is non-zero although quite small, often less than 10 bp. The spread will reflect three major factors. Supply-demand effects that prevail at t_0 , any transaction costs, any differences in credit ratings of the swap counterparties in case the swap involves currencies other than the major exchange rates.

I.B.4.2.5 Basis Swaps

Currency swaps are actually an example of the basis swap category, except that in case of a basis swap in general there will be only one currency involved. We say that an interest rate swap is a basis swap if the swap involves exchanging a cash flow in one *floating* rate to another *floating* rate in the same currency. Also, the principals will not be exchanged. One of the involved interest rates is a non-Libor based rate.

Example I.B.4.6 *Fannie Mae, a US government Agency borrows from international money markets in USD Libor and then lends these funds to mortgage banks at the US discount rate. Fannie Mae faces a basis risk in doing this. There is always a small difference between the interest rate that is paid, which is USD Libor, and the interest rate it receives, the USD discount rate. To hedge its position Fannie Mae needs to convert one floating rate to the other.*

Naturally, the pricing of basis swaps will be similar to the pricing of currency swaps except that there will be no exchange rate involved. However, because basis swaps deal with different types of interest rates that may be defined under different conventions and may be based on completely different credit exposures, the basis swap spread can vary a great deal and need not be small. For example, a basis swap between the prime rate and USD Libor will involve a spread around -250 basis points.

I.B.4.2.6 Volatility swaps

In a volatility swap a client sells floating volatility, and the market maker agrees to pay a fixed *volatility* on an agreed notional amount for a certain period. The floating volatility can for example be the annualized realized volatility for the S&P 500 for the life of the swap. Thus, the volatility or variance swaps are in many ways, just like any other swap. One pays (receives) depending on a *floating* risk, and receives (pays) depending on a risk fixed at the contract's origin. In this case however, what is being swapped is not an interest rate or a return on some equity instrument, but the volatilities that correspond to these risk factors.

I.B.4.3 Engineering Interest Rate Swaps

We now study how a swap can be re-constructed using simple cash flows or instruments. That is to say we look at the financial engineering of swaps. We focus on plain vanilla *interest rate* swaps. Engineering of other swaps is similar in many ways and is left to the reader. For simplicity, we deal with a simple case of 3 settlement dates. Figure I.B.4.5 shows a *fixed-payer*, 3-period interest rate swap, with start date t_o . The swap will be signed at date t_o . The counterparty will make 3 fixed payments and will make 3 floating payments at dates t_1, t_2, t_3 in the same currency. The dates t_1, t_2, t_3 are *settlement* dates and t_0, t_1, t_2 are the *reset* dates. The latter are dates on which the relevant Libor rate will be determined. We select the notional amount N as unity, and let $\delta = 1$, assuming that floating rate is 12-month Libor:²

$$N = \$1 \tag{I.B.4.3}$$

The fixed payments will be denoted by s_{t_o} and the Libor-linked payments will be $L_{t_o}N\delta, L_{t_1}N\delta, L_{t_2}N\delta$, respectively. The *swap spread* will be the difference between s_{t_o} and the benchmark risk-free rate corresponding to the same maturity, denoted by y_{t_o} ³. Thus we have:

$$\text{swap spread} = s_{t_o} - y_{t_o} \tag{I.B.4.4}$$

Such a swap can be reverse-engineered in different ways. We consider two interesting avenues.

1. One can first decompose the swap *horizontally*, into two streams of cash flows, one representing a floating stream of payments (receipts), the other a fixed stream. If this is done, then each stream can be interpreted as representing a certain type of bond. The fixed leg cash flows will correspond to a fixed coupon bond with coupon equal to the swap rate, and the floating leg will correspond to an FRN. Then the pricing proceeds by pricing each bond separately. The value of swap spread will equate the value of two bonds.
2. Second, one can decompose the swap *vertically*, slicing it into n cash exchanges during n time periods. If this is done, then each cash exchange can be priced using appropriate FRA-in-arrears.

Here we chose the first method since this is the one found in most textbooks, and hence is more widely used. We consider a *forward* swap that is signed at time t_o , which starts at time t_1 , with $t_o < t_1$. The traditional way to decompose an interest rate swap is to do this horizontally. The original swap cash flows are shown in the top part of Figure I.B.4.5.

We first use a trick. We add and subtract the same notional amount N at the start and end dates respectively for both sets of cash flows. Since these involve identical currencies and identical amounts, they cancel out and we recover the standard exchanges of floating versus fixed rate payments. But with the addition and subtraction of the initial principals, the swap will look as in Figure I.B.4.5.

²This is a simplification. In reality floating rate is either 3-month or 6-month Libor.

³This could be any interest rate accepted as a benchmark by the market.

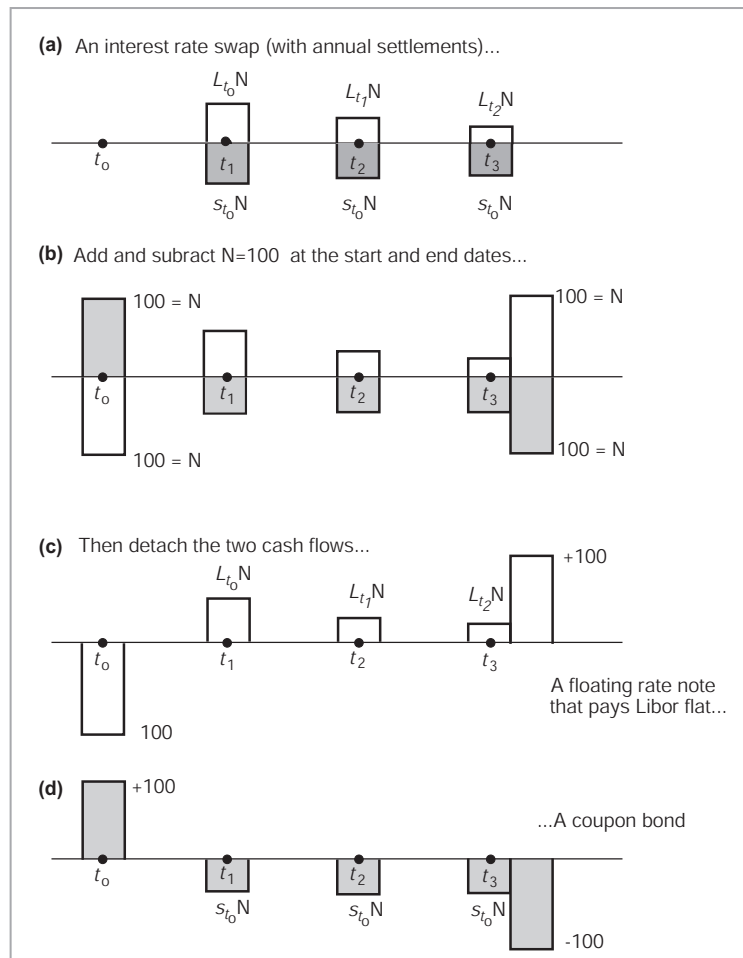


Figure I.B.4.5: Engineering a Swap.

We then ‘detach’ the cash flows in Figure I.B.4.5 horizontally, so as to obtain two separate cash flows. Note that each sequence of cash flows is already in the form of a meaningful financial contract. In fact, the cash flows in the middle part of Figure I.B.4.5 can immediately be recognized as a long forward position in a floating rate note that pays *Libor flat*. At time t_0 , the initial amount N is paid and L_{t_0} is set. At t_1 , the first interest payment is received and this will continue until time t_3 where the last interest is received along with the principal.

The remaining cash flows, on the other hand, can be recognized as a short forward position on a par coupon bond that pays a coupon equal to s_{t_0} . One (short) sells the bond to receive N . At every payment date the fixed coupon is paid and then, at t_3 , one pays back the last coupon and the principal N .

We now discuss pricing of interest rate swaps. The representation obtained above permits pricing the swap off the *debt markets*, using prices of fixed and floating coupon bonds. In order to obtain the present value of the fixed cash flows, we discount these cash flows by the relevant floating rates as

follows. For the fixed payments, if we knew all the $\{L_{t_0}, L_{t_1}, L_{t_2}\}$ we could write:

$$\text{PV-fixed} = \frac{s_{t_0} \delta N}{(1 + L_{t_0} \delta)} + \frac{s_{t_0} \delta N}{(1 + L_{t_0} \delta)(1 + L_{t_1} \delta)} + \frac{s_{t_0} \delta N + N}{(1 + L_{t_0} \delta)(1 + L_{t_1} \delta)(1 + L_{t_2} \delta)} \quad (\text{I.B.4.5})$$

But at $t = 0$, we don't know the two Libor rates $L_{t_i}, i = 1, 2$. Yet we know that against each $L_{t_i}, i = 1, 2$, the market is willing to pay the known forward (FRA) rate $F(t_0, t_i)$. Thus, using the FRA rates as if they are the time t_0 market values of the unknown Libor rates, we get:

$$\text{PV-fixed at } t_0 = \frac{s_{t_0} \delta N}{(1 + F(t_0, t_0) \delta)} + \frac{s_{t_0} \delta N}{(1 + F(t_0, t_0) \delta)(1 + F(t_0, t_1) \delta)} + \quad (\text{I.B.4.6})$$

$$\frac{s_{t_0} \delta N + N}{(1 + F(t_0, t_0) \delta)(1 + F(t_0, t_1) \delta)(1 + F(t_0, t_2) \delta)} \quad (\text{I.B.4.7})$$

All the right-hand side quantities are known and this present value can be calculated exactly, given the s_{t_0} . For the floating rate cash flows we have:

$$\text{PV-floating at } t_0 = \frac{L_{t_0} \delta N}{(1 + L_{t_0} \delta)} + \frac{L_{t_1} \delta N}{(1 + L_{t_0} \delta)(1 + L_{t_1} \delta)} + \frac{L_{t_2} \delta N + N}{(1 + L_{t_0} \delta)(1 + L_{t_1} \delta)(1 + L_{t_2} \delta)} \quad (\text{I.B.4.8})$$

Here to get a numerical answer, we don't even need to use the forward rates. This present value can be written in a much simpler fashion, as discussed above:

$$\text{PV of floating payments at } t_0 = N \quad (\text{I.B.4.9})$$

This means that the present value of a FRN is equal to the par value N at every settlement date. We can now combine these two present values and use the fact that:

$$\begin{aligned} & \frac{s_{t_0} \delta N}{(1 + F(t_0, t_0) \delta)} + \frac{s_{t_0} \delta N}{(1 + F(t_0, t_0) \delta)(1 + F(t_0, t_1) \delta)} \\ & + \frac{s_{t_0} \delta N + N}{(1 + F(t_0, t_0) \delta)(1 + F(t_0, t_1) \delta)(1 + F(t_0, t_2) \delta)} = N \end{aligned} \quad (\text{I.B.4.10})$$

Cancelling the N , the δ and rearranging, we can obtain the numerical value of s_{t_0} given $F(t_0, t_0), F(t_0, t_1), F(t_0, t_2)$. This would value the swap off the FRA markets.

$$B(t_0, t_1)[F(t_0, t_0) - s_{t_0}] \delta N + B(t_0, t_2)[F(t_0, t_1) - s_{t_0}] \delta N + B(t_0, t_3)[F(t_0, t_2) - s_{t_0}] \delta N = 0 \quad (\text{I.B.4.11})$$

Rearranging further we get a formula that ties the swap rate to FRA-rates:

$$s_{t_0} = \frac{B(t_0, t_1)F(t_0, t_0) + B(t_0, t_2)F(t_0, t_1) + B(t_0, t_3)F(t_0, t_2)}{B(t_0, t_1) + B(t_0, t_2) + B(t_0, t_3)} \quad (\text{I.B.4.12})$$

This means that we can price swaps relative to bond market as well. The general formula, where the n is the maturity of the swap will be given by

$$s_{t_0} = \frac{\sum_{i=1}^n B(t_0, t_i) F(t_0, t_{i-1})}{\sum_{i=1}^n B(t_0, t_i)} \quad (\text{I.B.4.13})$$

This gives an important arbitrage relationship

I.B.4.4 Risks of swaps

Let us briefly review the main risks involved in swaps.

I.B.4.4.1 Market risk

A typical swap consists of two legs, one fixed, the other floating. The risks of these two component will naturally differ. Newcomers to market finance may think that the risky component is the floating leg, since the underlying interest rate floats, and hence, is unknown. This first impression is wrong. The risky component is in fact the fixed leg and it is very easy to see why this is so.

The discussion of pricing interest rate swaps illustrated an important point. Regardless of what happens to future Libor rates, the value of a rolling deposit or FRN always equals the notional amount N at the reset dates. Between the reset dates this value may be different than N , but the discrepancy cannot be very large since the δ will be 3 or 6 months. Interest rate fluctuations have minimal effect on the values of fixed instruments with such maturities. In other words, the value of the floating leg changes very little during the life of a swap.

On the other hand the fixed leg of a swap is equivalent to a coupon bond and fluctuations of the swap rate may have major effects on the value of the future fixed payments.

I.B.4.4.2 Credit risk and counterparty risk

In general swaps will have no credit risk. Often the underlying principal is not exchanged. Hence swap parties do not advance any capital. In case the principals are exchanged, a principal of equal value is received for what is being advanced at time t_o . Thus, the default of one of the parties will involve no credit risk.

On the other hand, swaps may have counterparty risk. If the market has moved against one of the swap parties, the party that is negatively affected may default on future payments. This will lead to a loss of a capital gain. Yet, according to current practice if a retail client starts suffering from mark to market losses, the swap dealer asks for appropriate collateral. This reduces the counterparty risk as well.

I.B.4.4.3 Volatility and correlation risk

Some swaps have significant convexities and their value depends on the volatilities and correlations of the underlying forward rates significantly. One example is the Constant Maturity Swaps. When volatilities and correlations move, the value of such swaps may change and one of the parties may suffer mark-to-market losses.

I.B.4.5 Other Swaps

There are some additional terms and instruments that we would like to introduce before moving on.

A *FRA strips* is a sequence of FRA's whose settlements match the payment dates of an interest rate swap. A *futures strip* would be the similar idea for a commodity swap.

A *par swap* is a swap and the formal name of the interest rate swaps we have been showing on our figures in this chapter. It is basically denoted by a swap structure calculated over an initial and final (nominal) exchange of a principal equal to 100. This way there will be no additional cash payments at the time of signature. If, on the other hand, one buys, say, a (nominal) 95 at the initial date and then receives a (nominal) 100, then this would not be a par swap.

An *accrual swap* is an interest rate swap where one counterparty pays a standard floating reference rate, such as Libor, and receives Libor plus a spread. But the interest payments to the second counterparty will accrue only for days for which Libor stay within pre-set upper and lower boundaries.

Commodity-linked interest rate swap is a hybrid swap in which Libor is exchanged for a fixed rate linked to a commodity price. A buyer of crude oil may wish to tie his costs to the cost of his debt. He could elect to receive Libor and pay a crude oil-linked rate such that as the price of crude oil rises, the fixed rate he pays declines.

Crack spread swap is a swap used by oil refiners. They pay the floating price of the refined product, and receive the floating price of crude oil plus a fixed margin, the crack spread. This way refiners can hedge a narrowing of the spread between crude oil prices and the price of their refined products.

Overnight Index Swap This is an interest rate swap which uses some *index of overnight interest rates* for the floating leg. The interest is compounded and paid at settlement dates in exchange for the fixed rate.

A *Power Libor swap* is a swap that pays Libor squared or cubed (and so on) less a fixed amount/rate in exchange for a floating rate.

Extendible swap is a swap in which one counterparty has the right to extend a swap beyond its original term.

I.B.4.6 Uses of Swaps

Swaps can be used in *balance sheet management*. A corporation's balance sheet may contain one cash flow; using the swap corporation can switch cash flow characteristics. Swap will be used in *hedging*. They have *zero* value at time of initiation and hence they don't require any funding. A market practitioner can easily cover its positions in Equity, commodities and Fixed Income by quickly arranging proper swaps and then unwinding these positions when there is no need for the hedge. Finally, swap are *trading* instruments. In fact, one can construct *spread trades* most conveniently by using swaps. Some possible spreads trades are given by the following.

I.B.4.6.1 Uses of equity swaps

It is interesting that equity swaps are a better way of introducing versatility of instruments swaps, when compared with the much broader category of interest swaps. There is a huge industry of fund management where the fund manager tries to track some equity *index*. One way to do this is to buy the portfolio of stocks that replicates the index and constantly readjust it, as the market moves, or as new funds are received or paid by the fund. This requires a fairly complex operation. Of course, one can use S&P 500 futures to accomplish the same thing. But futures contracts need to be rolled over and they have mark-to-market adjustments. Equity swaps is a much more cost effective way of doing this. The fund manager could get into an equity swap on S&P500 swap in which the fund will pay, quarterly, a Libor related rate *and* a (positive or negative) spread and receive the return on S&P 500 index for a period on n years.

Equity swaps will, not only be cheaper and more efficient, but may have some *tax* and *ownership* advantages as well. For example, if an investor wants to sell a stock that appreciated significantly, then doing this through an outright sale will be subject to capital gains taxes. Instead the investor can keep the stock but get in an equity swap where he pays the capital gains (losses) and dividends and receives some Libor related return and a spread. Finally equity swaps make possible some strategies that otherwise were not possible due to some regulations.

Duration is the ‘average’ maturity of a fixed income portfolio. It turns out that often the largest fixed income liabilities are managed by governments, due to the existence of government debt. Depending on market conditions governments may want to adjust the average maturity of their debt. Swaps are very useful here.

I.B.4.7 Swap Conventions

Interest rate swap markets have their own conventions. The first question is the same: What do we quote? In some economies, the market quotes the swap *spread*. This is the case for USD interest rate swaps. USD interest rates swaps are quoted as a spread to treasuries. In Australia, the market also quotes swap spreads. But the spreads this time are with bond futures. In other economies, the market quotes the swap *rate*. This is the case in Euro interest rate swaps. This latter practice is regarded as creating more work from the point of view of market professionals. When only the spread is quoted, the positions are taken with respect to the spread and the hedging and risk management will be done only with respect to the spread. But with swap rate quotations, the hedging should involve the hedging for the spread, and the hedging for the underlying treasury rate. The market professionals have to hedge against more random factors in this latter case.

Next, there is the issue of how to quote swaps. This will be done in terms of two-way interest rate quotes. But sometimes the quoted swap rate is on an annual basis, and sometimes on a semi-annual

basis. Also, the day count changes from one market to another. In USD swaps, the day count is in general 30/360.

According to market conventions, a fixed payer is called *a payer*, is *long* the swap, and has *bought* a swap. On the other hand, a fixed receiver is called *a receiver*, is *short* the swap, and has *sold* a swap.

I.B.4.8 Conclusions

Swaps are a critical component of modern market practices. Their value at initiation is zero and this increases their liquidity. Swaps carry no credit risk and have limited counterparty risk. On the other hand from mortgage industry to credit derivatives, modern financial practices cannot be implemented without using swaps.

I.B.5 Vanilla Options

Paul Wilmott

I.B.5.1 Stock Options – Characteristics and Payoff Diagrams

The two most common types of option on equities are calls and puts. The European call option is the right to buy a specified asset at some specified time in the future for a specified amount. The put option is the right to sell a specified asset at some specified time. If you own a call option you want the asset to rise as much as possible so that you can buy the stock for a relatively small amount, then sell it and make a profit. If you own a put option you want the stock to fall in value.

Here is some option jargon.

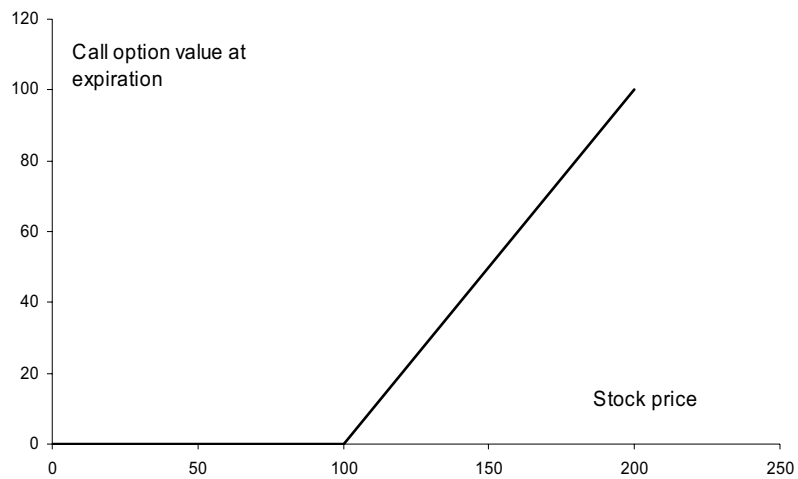
- *Premium*: The amount paid for the contract initially.
- *Underlying asset*: The financial instrument on which the option value depends. It could be a stock, commodity, currency, index, interest rate, etc.
- *Strike price* or *exercise price*: The amount for which the underlying can be bought (call) or sold (put).
- *Expiration date* or *expiry*: Date on which the option can be exercised or date on which the option ceases to exist or give the holder any rights.
- *In the money*: An option with positive intrinsic value (see Chapter 1.A.8) – a call option when the asset price is above the strike, a put option when the asset price is below the strike.
- *Out of the money*: An option with no intrinsic value, only time value – a call option when the asset price is below the strike, a put option when the asset price is above the strike.
- *At the money*: A call or put with a strike that is close to the current asset level.
- *At-the-money forward*: As ‘at the money’ but referring to the forward price of the stock rather than the spot price.
- *Long position*: A positive amount of a quantity, or a positive exposure to a quantity.
- *Short position*: A negative amount of a quantity, or a negative exposure to a quantity.

Understanding options is helped enormously by drawing diagrams to explain what an option is worth at expiration. These diagrams are called payoff diagrams.

If you own a call option and at expiration the stock is below the strike, the option is worthless. There is no point in paying the strike to get the asset when you could more cheaply buy the

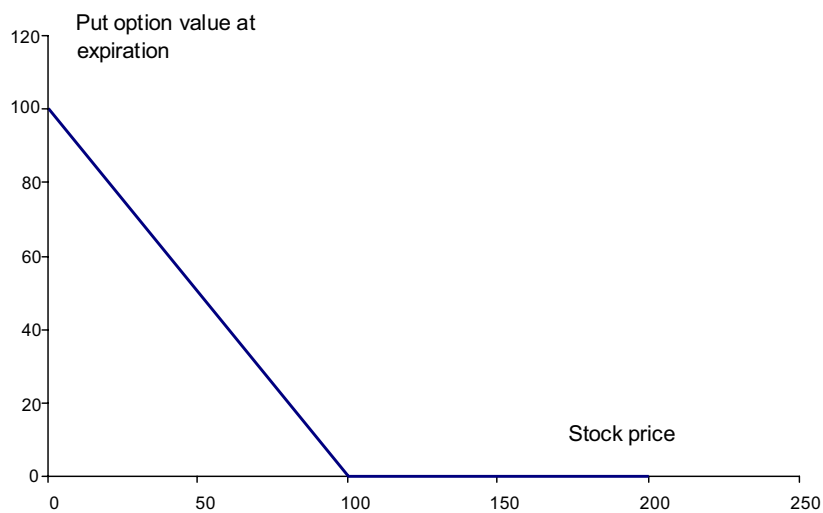
option directly. If the asset is above the strike at expiration, you profit by the difference between the stock price and the strike. Thus the payoff diagram is as shown in Figure I.B.5.1. The profit is realised here by purchasing the asset at the strike (from the option counterparty) and subsequently selling it at market prices.

Figure I.B.5.1



Similarly, if you hold a put option, it is worthless if the stock is above the strike. The payoff diagram is shown in Figure I.B.5.2. If the stock price is below the strike, a profit can be realised by purchasing the asset at market prices and selling it at the strike price to the option counterparty.

Figure I.B.5.2



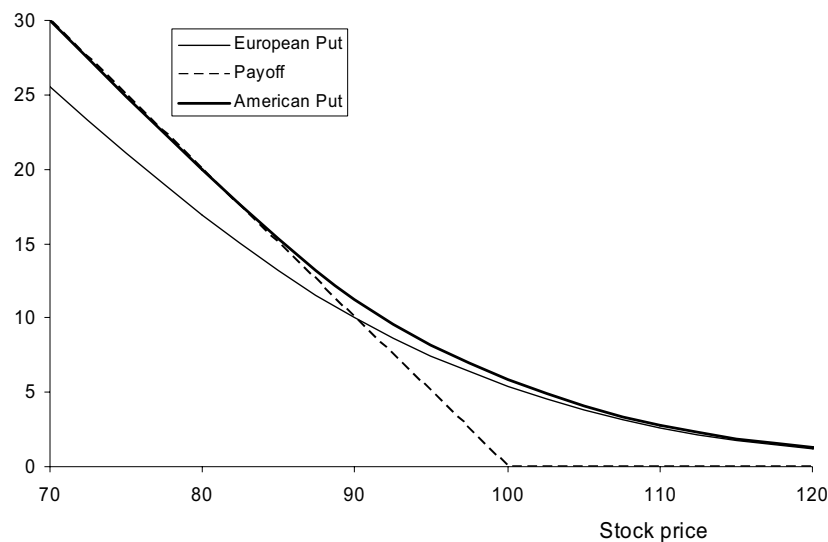
The above plots and descriptions relate to long positions, that is, a positive quantity of options. To understand short positions (written options) just reflect the lines in the stock-price axis. The key thing to note about short positions is that, once you have written an option and taken in the premium, things can only get worse; the best that can happen is that the option expires out of the money and you get to keep the premium. The downside for written options can be very large, unlike the downside for long positions, which is limited to the initial premium.

I.B.5.2 American versus European Options

American options are contracts that may be exercised early, prior to expiry. This is in contrast to European options, which may be exercised only at expiry. For example, if the option is an American call, we may hand over the exercise price and receive the asset whenever we wish. Most traded stock and futures options are American style, but most index options are European.

The right to exercise at any time at will is clearly valuable. The value of an American option cannot be less than an equivalent European option. But as well as giving the holder more rights, they also give him more headaches. When is the best time to exercise? Part of the valuation problem is deciding just that. This is what makes American options much more interesting than their European cousins. Figure I.B.5.3 shows the theoretical values of European and American puts (and their payoff). Note that the American put value is always at least as great as the payoff and always greater in value than the European put.

Figure I.B.5.3



Note that it is possible for European options to be worth less than the payoff. This is obvious for European puts when the stock price is low, but is also true for European calls when there are

dividends on the underlying and the stock is high in value. When the stock price is very high the call is almost certain to be exercised, so you would be better off holding the stock and getting the dividends than holding the option.

Bermudan options are a variant of American options. With these options there are specified dates or periods on which they may be exercised, and on other dates they cannot be. Typically Bermudan options will have a value not less than the equivalent European option and not greater than the American option. More details of Bermudan, American and other complex options are given in the Chapter on Exotic Options (Chapter I.B.9).

I.B.5.3 Strategies Involving a Single Option and a Stock

A common and very basic use of options is to generate some extra income from a stockholding. Imagine that you hold some stock, a stock that is perhaps not going anywhere fast. If you now write an option on that stock you are doing what is known as *covered call* writing. It is ‘covered’ because you already own the stock: if the option is exercised then you just hand over the stock. By writing this option you gain the premium. You will not lose from this position, the worst that could happen is that the stock price rises and you have missed out the profits you would otherwise have made. If, on the other hand, the stock price falls you have taken in the premium.

Example I.B.5.1: Covered call writing

A fund manager owns 1000 units of a stock, each currently valued at \$10.00. She expects that over the next six months the stock will not increase in value beyond \$11.00. In order to generate a higher return on the portfolio she sells 1000 European call options with a strike of \$11.00 and a term of six months. The premium income she receives today is \$0.2744 per option (see Example I.A.8.2).

Suppose that at expiry the stock price remains at \$10.00. In this case the option will not be exercised, but she will benefit from the premium income. Assuming that she invested her premium income at the risk-free rate, it has now grown in value to $1000 \times 0.2744e^{0.0392 \times 0.5} = \279.83 , which, combined with her original shareholding, gives a total portfolio value of $\$279.83 + \$10,000 = \$10,279.83$.

Suppose that at expiry the stock price increases to \$12.00. That is, her view is mistaken. In this case the option is exercised and she is required to sell her 1000 units of stock at the strike price of \$11.00. Now the total value of her portfolio (held in cash) is $\$279.83 + (\$11.00 \times 1000) = \$11,279.83$. In contrast, if she had not sold the calls she would have had 1000 shares valued at \$12.00 or a portfolio valued at \$12,000.

This example illustrates that selling covered calls has the effect of ‘giving away the upside’. That is, the option seller cannot benefit from increases in the share price beyond the strike. In addition, it does not provide any protection against a fall in the share price beyond the option premium that is earned.

Another simple strategy using a single-option position is the buying of puts on a stock that you already hold. This is called a *protective put* because the strategy offers protection, usually short-term, in case the stock price falls. The use of puts in this way is identical in spirit to the purchase of household insurance.

Example I.B.5.2: Protective puts

A different fund manager holds 1000 units of stock currently valued at \$10.00. He is fearful that the value of the stock will fall over the next six months. He purchases 1000 put options with a strike price of \$10.00 and a term of six months at a cost of \$0.4665 per option.

Suppose that at expiry the share price falls to \$9.00. He can exercise his right to sell the shares at \$10.00, which will generate cash of \$10,000.00. The value of the option protection in today’s dollars is $1000 \times 0.4665e^{0.0392 \times 0.5} = \475.73 . The net value of his position is therefore $\$10,000.00 - \475.73 or \$9,524.27. Had he not purchased the put options the value of his shares would have fallen to \$9,000.00.

What if the share price instead increases to \$11.00? In this case, his view is mistaken. The put options are not exercised so he remains holding shares worth \$11,000.00. After taking account of the option premium, the value of the position is $\$11,000.00 - \475.73 or \$10,524.27. The option premium is a drag on portfolio performance; without it the portfolio would be worth \$11,000.00.

I.B.5.4 Spread Strategies

A spread of options is a portfolio containing options of different types (calls and puts) and/or different strikes. They enable one to benefit from precise (if correct) views on the movement of the stock.

I.B.5.4.1 Bull and Bear Spreads

Suppose I buy one call option with a strike of 100 and write another with a strike of 120 and with the same expiration as the first. Then my resulting portfolio has a payoff that is shown in Figure I.B.5.4. This payoff is zero below 100, 20 above 120 and linear in between. This strategy is called a bull spread (or a call spread) because it benefits from a bull (i.e. rising) market. If I write a put

option with strike 100 and buy a put with strike 120, I get the payoff shown in Figure I.B.5.5. This is called a bear spread (or put spread), benefiting from a bear (i.e. falling) market.

Because of put–call parity it is possible to build up these payoffs using other contracts.

Figure I.B.5.4

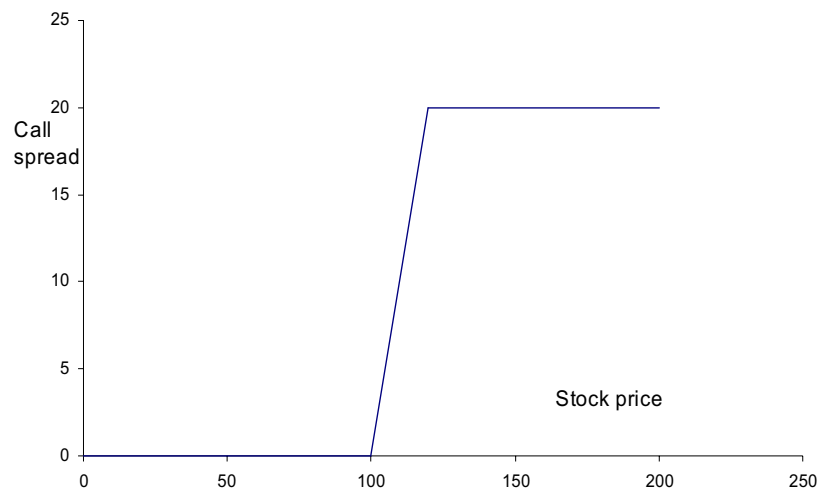
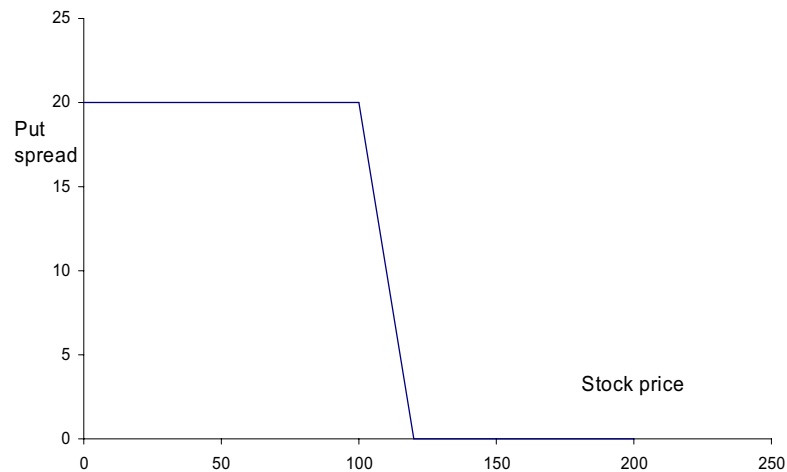


Figure I.B.5.5



I.B.5.4.2 Calendar Spreads

The strategies described above involve buying or writing calls and puts with different strikes, but all with the same expiration. A strategy involving options with different expiry dates is called a calendar spread. You may enter into such a position if you have a precise view on the timing of a market move as well as the direction of the move. As always, the motive behind such a strategy is

to reduce the payoff at asset values and times that you believe are irrelevant, while increasing the payoff where you think it will matter. Any reduction in payoff will reduce the overall value of the option position.

I.B.5.5 Other Strategies

I.B.5.5.1 Straddles and Strangles

If you have a precise view on the behaviour of the underlying asset you may want to be precise in your choice of option. Simple calls and puts may not have the precise payoff that you require.

The straddle consists of a call and a put with the same strike. The payoff diagram is shown in Figure I.B.5.6. Such a position is usually bought at the money by someone who expects the underlying price to either rise or fall, but not to remain at the same level. For example, just before an anticipated major news item, stock prices often show a calm before the storm. On the announcement the stock suddenly moves either up or down in price, depending on whether or not the news was favourable to the company. Straddles may also be bought by technical traders who see the stock price at a key support or resistance level and expect the price to either break through dramatically or bounce back. The straddle would be sold by someone with the opposite view, someone who expects the underlying price to remain stable. Selling options can be a very risky business, because of the large downside exposure, as discovered by Leeson and others.

The strangle is similar to the straddle, except that the strikes of the put and the call are different. The motivation behind the purchase of this position is similar to that for the purchase of a straddle. The difference is that the buyer expects an even larger move in the underlying price one way or the other. The contract is usually bought when the asset price is around the middle of the two strikes and is cheaper than the corresponding straddle. This cheapness means that the gearing for the out-of-the-money strangle is higher than that for the straddle. The downside is that there is a much greater range over which the strangle has no payoff at expiry. For the straddle there is only the one point at which there is no payoff. This is clearly seen in Figure I.B.5.7.

There is another reason for a straddle or strangle trade that does not involve a view on the direction of the underlying. These contracts are bought or sold by those with a view on the direction of volatility: they are one of the simplest volatility trades. Because of the positive relationship between the price of an option and the volatility of the asset, one can use straddles and strangles to speculate on the direction of volatility.

I.B.5.5.2 Risk Reversal

The risk reversal is a combination of a long call, with strike above the current spot, and a short put, with a strike below the current spot. Both have the same expiry. The payoff is shown in Figure I.B.5.8.

Figure I.B.5.6

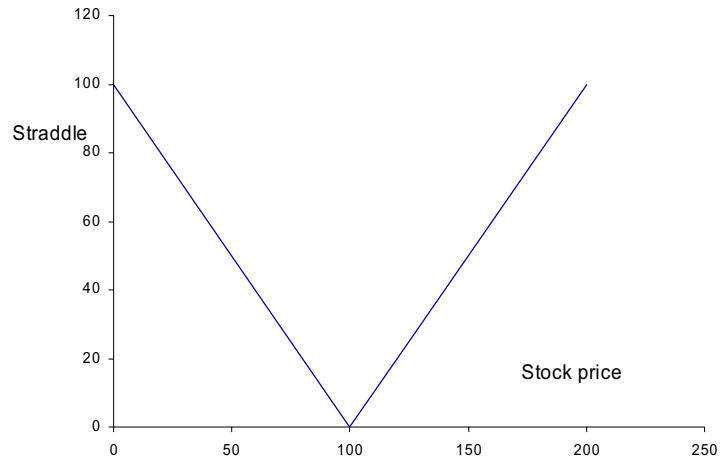


Figure I.B.5.7

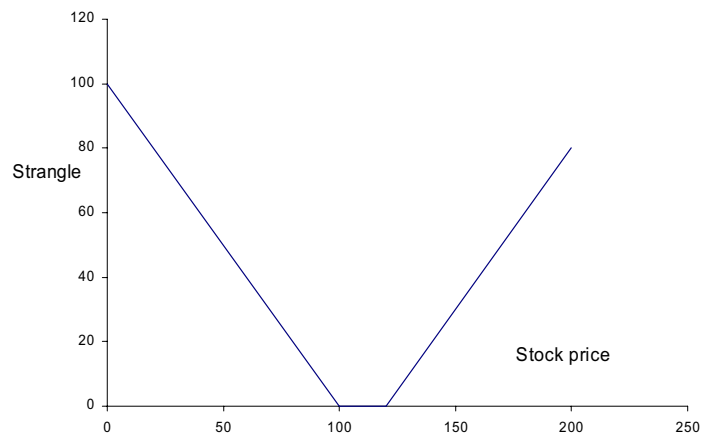
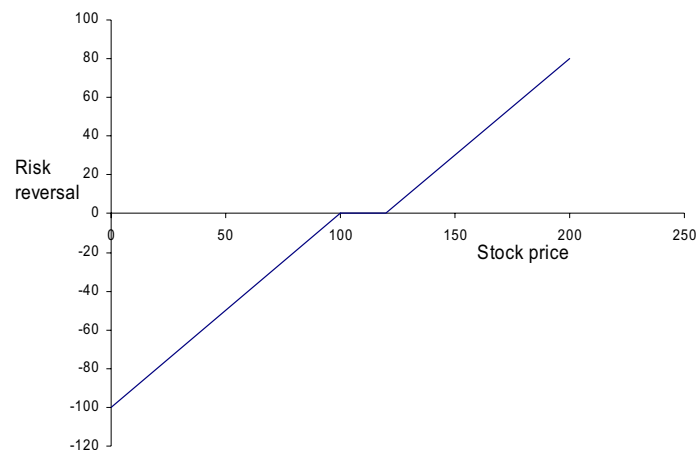


Figure I.B.5.8



The risk reversal is a very popular contract. Its value is usually quite small and related to the market's expectations of the behaviour of volatility. The value depends not so much on the absolute level of volatility but on the skew. A negative skew will give the risk reversal a low value, but a positive skew will give it a high value. Risk reversals are good if you want to bet on the behaviour of the skew.

I.B.5.5.3 Collars

This is a strategy involving the writing of a call and the purchase of a put while owning the underlying stock. The strikes of the two options would typically be different.

Example I.B.5.3: Collar strategy

Example I.B.5.2 shows that purchasing protective puts can create an expensive drag on portfolio performance. One way of alleviating this cost is to simultaneously sell call options at a higher strike price.

Consider yet another fund manager holding 1000 shares valued at \$10.00. To protect against a fall in their value she purchases 1000 put options with a strike of \$10.00, and a term of six months, costing \$0.4665 each. To offset this cost she also sells 1000 call options with a strike of \$11.00, a term of six months, receiving \$0.2744 per share. The net cost of the option strategy is therefore $\$0.4665 - \$0.2744 = \$0.1921$ per option. In future value terms this is a cost of $1000 \times 0.1921 e^{0.0392 \times 0.5} = \195.90 .

If the stock price at expiry is \$9.00, only the put options are exercised. The fund manager sells 1000 shares at \$10.00 each. After taking account of the option premium, this leaves cash of \$9804.10. This compares with a portfolio of \$9000.00 without protection.

If the stock price at expiry is \$10.00, neither option is exercised. The value of the position is \$10,000 less the premium, which is \$9804.10.

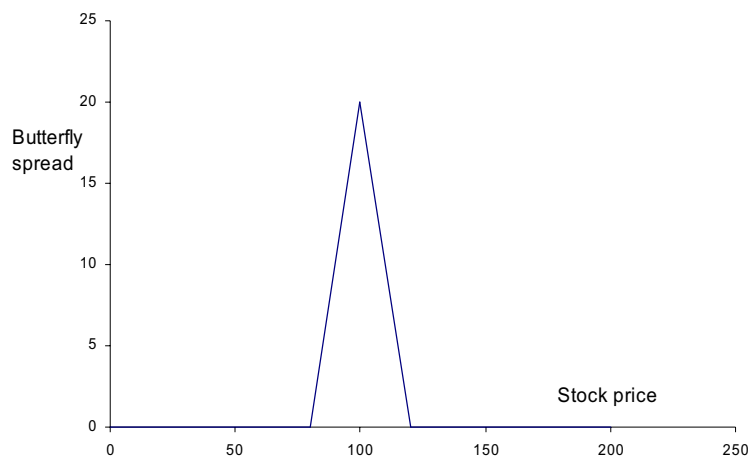
If the stock price at expiry is \$12.00 only the call options are exercised. The fund manager must sell 1000 shares at \$11.00 each. Her cash holdings are now $\$11,000 - \$195.90 = \$10,804.10$. Had she not entered into the option strategy at all she would have had shares worth \$12,000.00.

In summary, a collar strategy gives protection with limited upside. The cost is lower than that of purchasing protection in isolation.

I.B.5.5.4 Butterflies and Condors

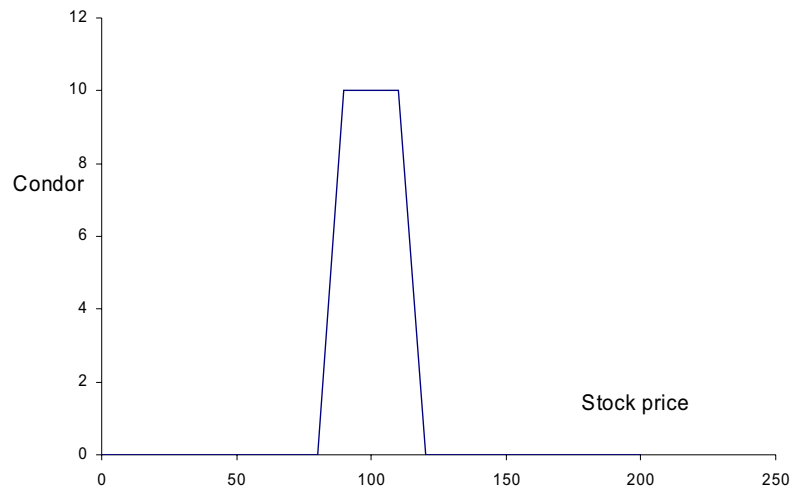
A more complicated strategy involving the purchase and sale of options with three different strikes is a *butterfly spread*. Buying a call with a strike of 90, writing two calls struck at 100 and buying a 110 call gives the payoff in Figure I.B.5.9. This is the kind of position you might enter if you believe that the asset is not going anywhere, either up or down. Because it has no large upside potential (for example, in this case the maximum payoff is 10) the position will be relatively cheap.

Figure I.B.5.9



The *condor* is like a butterfly except that four strikes and four call options are used. The payoff is shown in Figure I.B.5.10. The condor has a larger range having a decent payoff, consequently it will cost more, than a similar butterfly.

Figure I.B.5.10



For other option strategies see Taleb, N (1997) *Dynamic Hedging* (New York: Wiley).

I.B.6 Credit Derivatives

Moorad Choudhry¹

This chapter describes credit derivatives, instruments that are used to manage credit risk in banking and portfolio management. In this chapter we consider only the most commonly encountered credit derivative instruments. Credit derivatives exist in a number of forms. We classify these into two main forms, *funded* and *unfunded* credit derivatives, and give a description of each form. We then discuss the main uses of these instruments by banks and portfolio managers. We also consider the main credit events that act as triggering events under which payouts are made on credit derivative contracts.

I.B.6.1 Introduction

Credit derivatives are financial contracts designed to hedge credit risk exposure by providing insurance against losses suffered due to credit events. Credit derivatives allow investors to manage the credit risk exposure of their portfolios or asset holdings, essentially by providing insurance against deterioration in credit quality of the borrowing entity. The simplest credit derivative works exactly like an insurance policy, with regular premiums paid by the protection buyer to the protection seller, and a payout in the event of a specified credit event

The principle behind credit derivatives is straightforward. Investors desire exposure to debt that has a risk of defaulting because of the higher returns this offers. However, such exposure brings with it concomitant credit risk. This can be managed with credit derivatives. At the same time, the exposure itself can be taken on synthetically if, for instance, there are compelling reasons why a cash market position cannot be established. The flexibility of credit derivatives provides users a number of advantages, and as they are over-the-counter products they can be designed to meet specific user requirements.

What constitutes a credit event is defined specifically in the legal documents that describe the credit derivative contract. A number of events may be defined as credit events that fall short of full bankruptcy, administration or liquidation of a company. For instance, credit derivatives contracts may be required to pay out under both technical as well as actual default.

A *technical default* is a delay in timely payment of an obligation, or a non-payment altogether. If an obligor misses a payment, by even one day, it is said to be in technical default. This delay may be

¹ Visiting Professor, Department of Economics, Finance and International Business, London Metropolitan University.

for operational reasons (and so not really a great worry) or it may reflect a short-term cash flow crisis, such as the Argentina debt default for three months. But if the obligor states it intends to pay the obligation as soon as it can, and specifies a time-span that is within (say) one to three months, then while it is in technical default it is not in actual default. If an obligor is in *actual default*, it is in default and declared as being in default. This does not mean a mere delay of payment. If an obligor does not pay, and does not declare an intention to pay an obligation, it may then be classified by the ratings agencies as being in 'default' and rated 'D'.

If there is a technical or actual default by the borrower so that, for instance, a bond is marked down in price, the losses suffered by the investor can be recouped in part or in full through the payout made by the credit derivative. A payout under a credit derivative is triggered by a *credit event*. As banks define default in different ways, the terms under which a credit derivative is executed usually include a specification of what constitutes a credit event.

I.B.6.1.1 Why Use Credit Derivatives?

Credit derivative instruments enable participants in the financial market to trade in credit as an asset, as they isolate and transfer credit risk. They also enable the market to separate funding considerations from credit risk.

Credit derivatives have two main types of application:

- *Diversifying the credit portfolio* A bank or portfolio manager may wish to take on credit exposure by providing credit protection in return for a fee. This enhances income on their portfolio. They may sell credit derivatives to enable non-financial counterparties to gain credit exposures, if these clients are unable or unwilling to purchase the assets directly. In this respect the bank or asset manager performs a credit intermediation role.
- *Reducing credit exposure* A bank can reduce credit exposure either for an individual loan or a sectoral concentration by buying a credit default swap. This may be desirable for assets that cannot be sold for client relationship or tax reasons. For fixed-income managers a particular asset or collection of assets may be viewed as an attractive holding in the long term, but at risk from short-term downward price movement. In this instance a sale would not fit in with long-term objectives; however, short-term credit protection can be obtained via a credit swap. For instance, a bank can buy credit protection on a BB-rated entity from a AA-rated bank. It then has eliminated its credit risk to the BB entity, and substituted it for AA-rated counterparty risk. Notice that as the bank retains a counterparty risk to the credit default swap issuer, one could argue that its credit risk exposure is never completely removed. In practice this is not a serious problem since the bank can manage counterparty risk through

careful selection and diversification of counterparties. In fact, in the interest-rate swap market, AA (interbank) quality is now considered a proxy for the government benchmark.

The intense competition amongst commercial banks, combined with rapid disintermediation, has meant that banks have been forced to evaluate their lending policy with a view to improving profitability and return on capital. The use of credit derivatives assists banks with restructuring their businesses, because they allow banks to repackage and parcel out credit risk, while retaining assets on balance sheet (when required) and thus maintaining client relationships. As the instruments isolate certain aspects of credit risk from the underlying loan or bond and transfer them to another entity, it becomes possible to separate the ownership and management of credit risk from the other features of ownership of the assets in question. This means that illiquid assets such as bank loans and illiquid bonds can have their credit risk exposures transferred; the bank owning the assets can protect against credit loss even if it cannot transfer the assets themselves.

The same principles carry over to the credit risk exposures of portfolio managers. For fixed-income portfolio managers some of the advantages of credit derivatives include the following:

- They can be tailor-made to meet the specific requirements of the entity buying the risk protection, as opposed to the liquidity or term of the underlying reference asset.
- They can be ‘sold short’ without risk of a liquidity or delivery squeeze, as it is a specific credit risk that is being traded. In the cash market it is not possible to ‘sell short’ a bank loan, for example, but a credit derivative can be used to establish synthetically the same economic effect.
- As they theoretically isolate credit risk from other factors such as client relationships and interest rate risk, credit derivatives introduce a formal pricing mechanism to price credit issues only. This means a market can develop in credit only, allowing more efficient pricing; it even becomes possible to model a term structure of credit rates.
- When credit derivatives are embedded in certain fixed-income products, such as structured notes and credit-linked notes, they are then off-balance-sheet instruments (albeit part of a structure that may have on-balance-sheet elements) and as such incorporate tremendous flexibility and leverage, exactly like other financial derivatives. For instance, bank loans are not particularly attractive investments for certain investors because of the administration required in managing and servicing a loan portfolio. However, an exposure to bank loans and their associated return can be achieved by a total return swap, for instance, while simultaneously avoiding the administrative costs of actually owning the assets. Hence, credit derivatives allow investors access to specific credits while allowing banks access to further distribution for bank loan credit risk.

- They enable institutions to take a view on credit positions to take advantage of perceived anomalies in the price of secondary market loans and bonds, and the price of credit risk.

Thus credit derivatives can be an important instrument for bond portfolio managers as well as commercial banks wishing to increase the liquidity of their portfolios, gain from the relative value arising from credit pricing anomalies, and enhance portfolio returns.

I.B.6.1.2 Classification of Credit Derivative Instruments

A number of instruments come under the category of credit derivatives. Irrespective of the particular instrument under consideration, all credit derivatives can be described with respect to the following characteristics:

- the *reference entity*, which is the asset or name on which credit protection is being bought and sold;
- the *credit event*, or events, which indicate that the reference entity is experiencing or about to experience financial difficulty and which act as trigger events for payments under the credit derivative contract;
- the *settlement mechanism* for the contract, whether cash settled or physically settled; and,
- the *deliverable obligation* that the protection buyer delivers (under physical settlement) to the protection seller on the occurrence of a trigger event.

Credit derivatives are grouped into *funded* and *unfunded* instruments. In a funded credit derivative, typified by a credit-linked note (CLN), the investor in the note is the credit-protection seller and is making an upfront payment to the protection buyer when it buys the note. Thus, the protection buyer is the issuer of the note. If no credit event occurs during the life of the note, the redemption value of the note is paid to the investor on maturity. If a credit event does occur, then on maturity a value less than par will be paid out to the investor. This value will be reduced by the nominal value of the reference asset that the CLN is linked to. The exact process will differ according to whether *cash settlement* or *physical settlement* has been specified for the note. We will consider this later.

In an unfunded credit derivative, typified by a credit default swap (CDS), the protection seller does not make an upfront payment to the protection buyer. Instead, the protection seller will pay the nominal value of the contract (the amount insured, in effect), on occurrence of a credit event, minus the current market value of the asset or its recovery value.

I.B.6.1.3 Definition of a Credit Event

The occurrence of a specified credit event will trigger the default payment by the protection seller to the protection buyer. Contracts specify physical or cash settlement. In physical settlement, the protection buyer transfers to the protection seller the deliverable obligation (usually the reference asset or assets), with the total principal outstanding equal to the nominal value specified in the default swap contract. The protection seller simultaneously pays to the buyer 100% of the nominal value. In cash settlement, the protection seller hands to the buyer the difference between the nominal amount of the default swap and the final value for the same nominal amount of the reference asset. This final value is usually determined by means of a poll of dealer banks.

The following may be specified as credit events in the legal documentation between counterparties:

- downgrade in S&P and/or Moody's and/or Fitch credit rating below a specified minimum level;
- financial or debt restructuring, for example occasioned under administration or as required under US bankruptcy protection;
- bankruptcy or insolvency of the reference asset obligor;
- default on payment obligations such as bond coupon and continued non-payment after a specified time period;
- technical default, for example the non-payment of interest or coupon when it falls due;
- a change in credit spread payable by the obligor above a specified maximum level.

The 1999 International Swaps and Derivatives Association (ISDA) credit default swap documentation specifies bankruptcy, failure to pay, obligation default, debt moratorium and 'restructuring' to be credit events. Note that it does not specify a rating downgrade to be a credit event.²

The precise definition of 'restructuring' is open to debate and has resulted in legal disputes between protection buyers and sellers. Prior to issuing its 1999 definitions, ISDA had specified restructuring as an event or events that resulted in making the terms of the reference obligation 'materially less favourable' to the creditor (or protection seller) from an economic perspective. This definition is open to more than one interpretation and caused controversy when determining if a credit event had occurred. The 2001 definitions specified more precise conditions, including any action that resulted in a reduction in the amount of principal. In the European market restructuring is generally retained as a credit event in contract documentation, but in the US market it is less common to see it included. Instead, US contract documentation

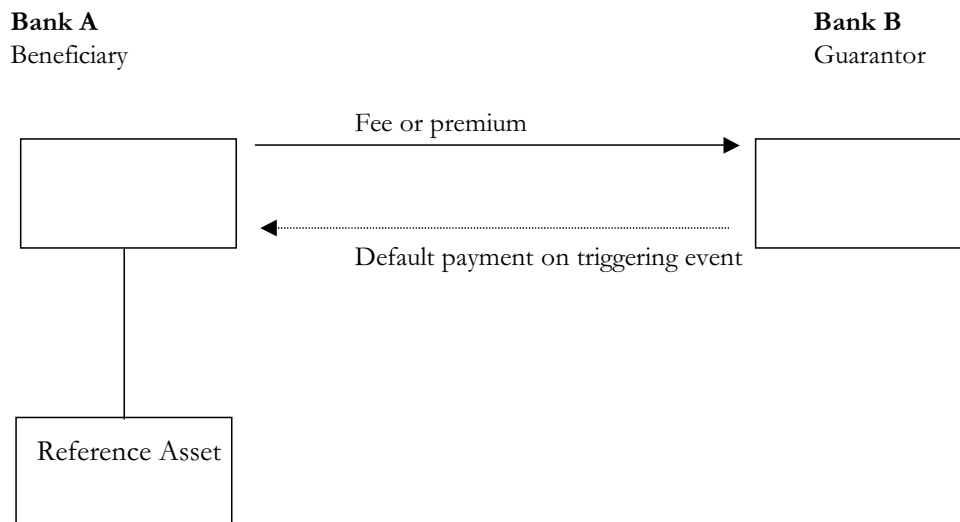
tends to include as a credit event a form of modified restructuring, the impact of which is to limit the options available to the protection buyer as to the type of assets it could deliver in a physically settled contract.

I.B.6.2 Credit Default Swaps

The most common credit derivative is the *credit default swap*. This is sometimes abbreviated to *credit swap* or *default swap*. A CDS is a bilateral contract in which a periodic fixed fee or a one-off premium is paid to a protection seller, in return for which the seller will make a payment on the occurrence of a specified credit event. The fee is usually quoted as a basis point multiplier of the nominal value. It is usually paid quarterly in arrears.

The swap can refer to a single asset, known as the *reference asset* or *underlying asset*, or a basket of assets. The default payment can be paid in whatever way suits the protection buyer or both counterparties. For example, it may be linked to the change in price of the reference asset or another specified asset, it may be fixed at a predetermined recovery rate, or it may be in the form of actual delivery of the reference asset at a specified price. The basic structure is illustrated in Figure I.B.6.1.

Figure I.B.6.1 Credit default swap



The maturity of the credit swap does not have to match the maturity of the reference asset and often does not. On occurrence of a credit event, the swap contract is terminated and a settlement

² The ISDA definitions from 1999 and restructuring supplement from 2001 are available at www.ISDA.org

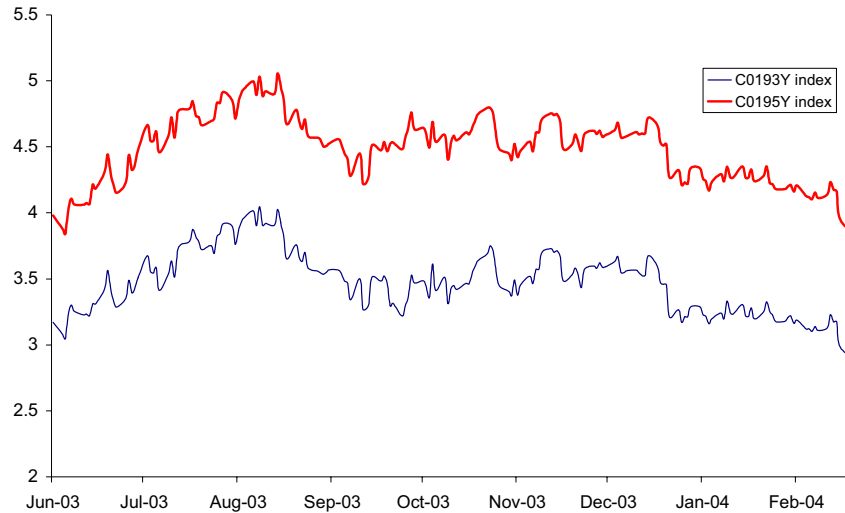
payment made by the protection seller or guarantor to the protection buyer. This termination value is calculated at the time of the credit event, and the exact procedure that is followed to calculate the termination value will depend on the settlement terms specified in the contract. This will be either cash settlement or physical settlement:

- *Cash settlement* The contract may specify a predetermined payout value on occurrence of a credit event. This may be the nominal value of the swap contract. Such a swap is known in some markets as a *digital credit derivative*. Alternatively, the termination payment is calculated as the difference between the nominal value of the reference asset and its market value at the time of the credit event. This arrangement is more common with cash-settled contracts.³
- *Physical settlement* On occurrence of a credit event, the buyer delivers the reference asset to the seller, in return for which the seller pays the face value of the delivered asset to the buyer. The contract may specify a number of alternative assets that the buyer can deliver; these are known as *deliverable obligations*. This may apply when a swap has been entered into on a reference name rather than a specific obligation (such as a particular bond) issued by that name. Where more than one deliverable obligation is specified, the protection buyer will invariably deliver the asset that is the cheapest on the list of eligible assets. This gives rise to the concept of the *cheapest to deliver*, as encountered with government bond futures contracts, and is in effect an embedded option afforded the protection buyer.

In theory, the value of protection is identical irrespective of which settlement option is selected. However, under physical settlement the protection seller can gain if there is a recovery value that can be extracted from the defaulted asset; or its value may rise as the fortunes of the issuer improve. Despite this, swap market-making banks often prefer cash settlement as there is less administration associated with it. It is also more suitable when the swap is used as part of a synthetic structured product, because such vehicles may not be set up to take delivery of physical assets. Another advantage of cash settlement is that it does not expose the protection buyer to any risks should there not be any deliverable assets in the market, for instance due to shortage of liquidity in the market. Were this to happen, the buyer may find the value of its settlement payment reduced. Nevertheless, physical settlement is widely used because counterparties wish to avoid the difficulties associated with determining the market value of the reference asset under cash settlement. Physical settlement also permits the protection seller to take part in the creditor negotiations with the reference entity's administrators, which may result in improved terms for them as holders of the asset.

Figure I.B.6.2 shows US dollar CDS price levels (in basis points) during 2003 and 2004 for BBB-rated reference entities, for three- and five-year CDS contracts. The graph shows the level of fluctuation in CDS prices, it also shows clearly the term structure of credit rates, as the five-year CDS price lies above the three-year rate at all times.

Figure I.B.6.2 Investment-grade credit default swap levels (Source: Bloomberg)



I.B.6.3 Credit-Linked Notes

A standard *credit-linked note* is a security, usually issued by an investment-grade entity, that has an interest payment and fixed maturity structure similar to a vanilla bond. The performance of the note, however, including the maturity value, is linked to the performance of a specified underlying asset or assets, as well as to that of the issuing entity. Notes are usually issued at par. The notes are often used by borrowers to hedge against credit risk, and by investors to enhance the yield received on their holdings. Hence, the issuer of the note is the protection buyer and the buyer of the note is the protection seller.

Credit-linked notes are essentially hybrid instruments that combine a credit derivative with a vanilla bond. The CLN pays regular coupons; however, the credit derivative element is usually set to allow the issuer to decrease the principal amount if a credit event occurs. For example, consider an issuer of credit cards that wants to fund its (credit card) loan portfolio via an issue of debt. In order to hedge the credit risk of the portfolio, it issues a two-year CLN. The principal amount of the bond is 100% as usual, and it pays a coupon of 7.50%, which is 200 basis points above the two-year benchmark. If, however, the incidence of bad debt amongst credit card holders exceeds 10% then the terms state that note holders will only receive back £85 per £100 nominal. The credit card issuer has in effect purchased a credit option that lowers its liability in

the even that it suffers from a specified credit event, which in this case is an above-expected incidence of bad debts. The credit card bank has issued the CLN to reduce its credit exposure, in the form of this particular type of credit insurance. If the incidence of bad debts is low, the note is redeemed at par. However, if there a high incidence of such debt, the bank will only have to repay a part of its loan liability.

Figure I.B.6.3 A cash-settled credit-linked note

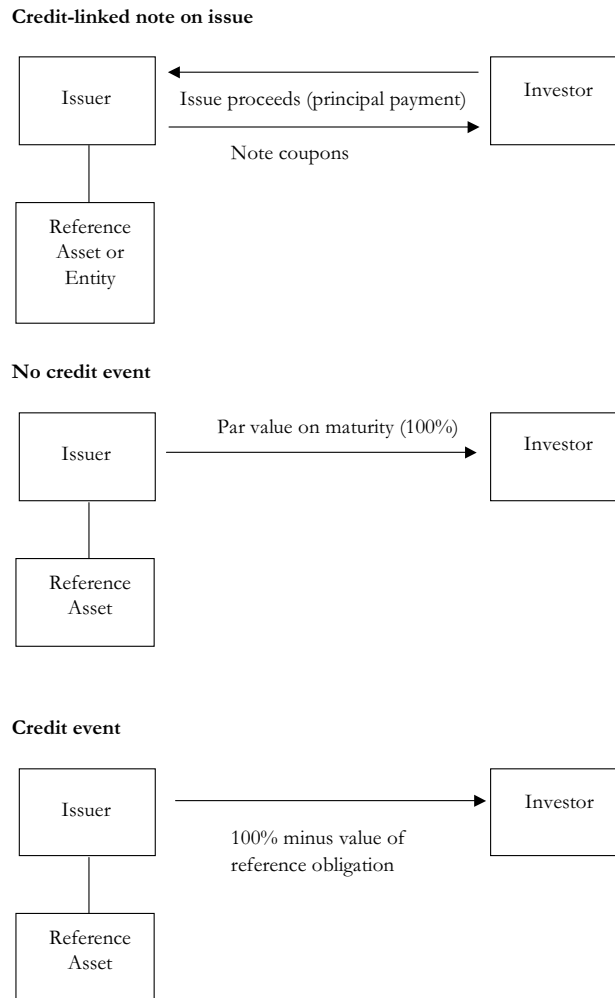
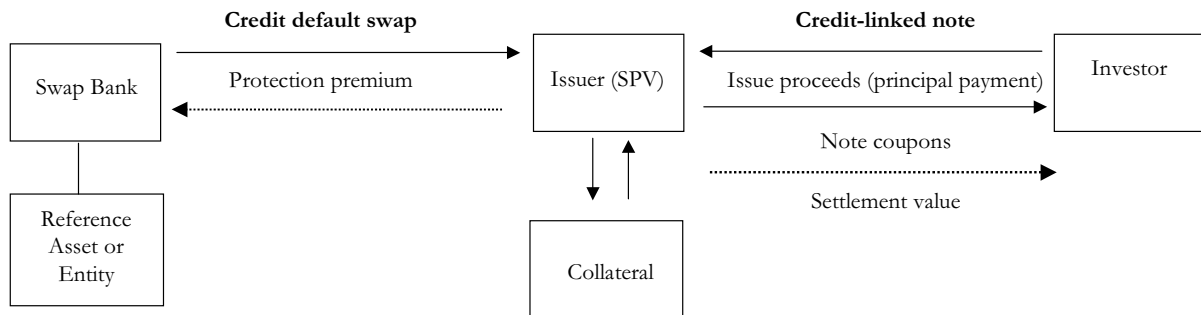


Figure I.B.6.3 depicts the cash flows associated with a credit-linked note. Credit-linked notes exist in a number of forms, but all of them contain a link between the return they pay and the credit-related performance of the underlying asset. Investors may wish to purchase the CLN because the coupon paid on it will be above what the same bank would pay on a vanilla bond it issued, and higher than other comparable investments in the market. In addition, such notes are usually priced below par on issue. Assuming the notes are eventually redeemed at par, investors will also have realised a substantial capital gain.

As with credit default swaps, CLNs may be specified under cash settlement or physical settlement. Specifically, under:

- *cash settlement*, if a credit event has occurred, on maturity the protection seller receives the difference between the value of the initial purchase proceeds and the value of the reference asset at the time of the credit event;
- *physical settlement*, on occurrence of a credit event, at maturity the protection buyer delivers the reference asset or an asset among a list of deliverable assets, and the protection seller receives the value of the original purchase proceeds minus the value of the asset that has been delivered.

Figure I.B.6.4 CLN and CDS structure on single reference name



Structured products may combine both CLNs and CDSs to meet issuer and investor requirements. For instance, Figure I.B.6.4 shows a credit structure designed to provide a higher return for an investor on comparable risk to the cash market. An issuing entity is set up in the form of a *special-purpose vehicle* (SPV) which issues CLNs to the market. The structure is engineered so that the SPV has a neutral position on a reference asset. It has bought protection on a single reference name by issuing a funded credit derivative, the CLN, and simultaneously sold protection on this name by selling a CDS on this name.

The proceeds of the CLN are invested in risk-free collateral such as T-bills or a Treasury bank account. The coupon on the CLN will be a spread over LIBOR. It is backed by the collateral account and the fee generated by the SPV in selling protection with the CDS. Investors in the CLN will have exposure to the reference asset or entity, and the repayment of the note is linked to the performance of the reference entity. If a credit event occurs, the maturity date of the CLN is brought forward and the note is settled at par minus the value of the reference asset or entity.

I.B.6.4 Total Return Swaps

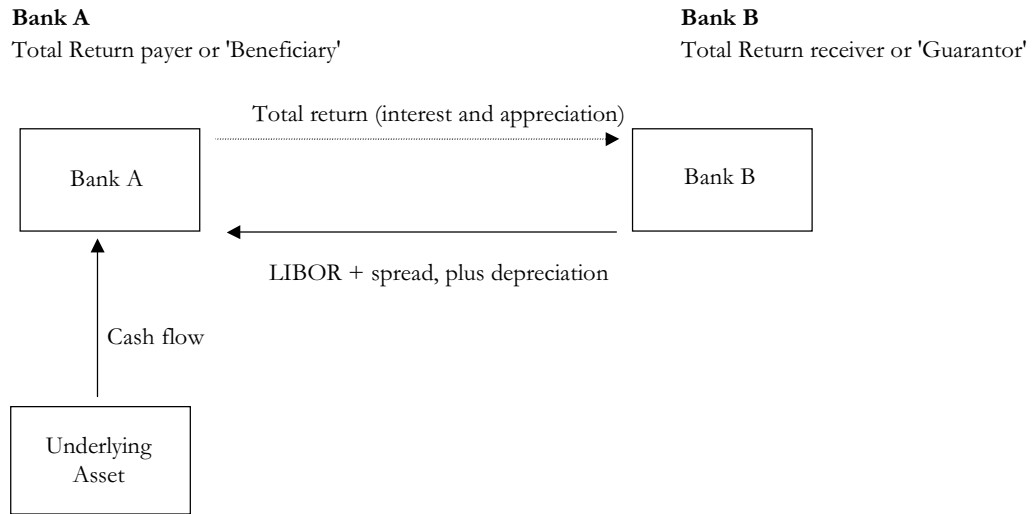
A *total return swap* (TRS), sometimes known as a *total rate of return swap* or *TR swap*, is an agreement between two parties to exchange the total returns from financial assets. This is designed to transfer the credit risk from one party to the other. It is one of the principal instruments used by banks and other financial instruments to manage their credit risk exposure, and as such is a credit derivative. One definition of a TRS is given in Francis *et al.* (1999), which states that a TRS is a swap agreement in which the *total return* of a bank loan or credit-sensitive security is exchanged for some other cash flow, usually tied to LIBOR or some other loan or credit-sensitive security.

In some versions of a TRS the actual underlying asset is actually sold to the counterparty, with a corresponding swap transaction agreed alongside; in other versions there is no physical change of ownership of the underlying asset. The TRS trade itself can be to any maturity term, that is, it need not match the maturity of the underlying security. In a TRS the total return from the underlying asset is paid over to the counterparty in return for a fixed or floating cash flow. This makes it slightly different from other credit derivatives, as the payments between counterparties to a TRS are connected to changes in the market value of the underlying asset, as well as changes resulting from the occurrence of a credit event.

Figure I.B.6.5 illustrates a generic TR swap. The two counterparties are labeled as banks, but the party termed 'Bank A' can be another financial institution, including insurance companies and hedge funds that often hold fixed income portfolios. In Figure I.B.6.5 Bank A has contracted to pay the 'total return' on a specified reference asset, while simultaneously receiving a LIBOR-based return from Bank B. The reference or underlying asset can be a bank loan such as a corporate loan or a sovereign or corporate bond. The total return payments from Bank A include the interest payments on the underlying loan as well as any appreciation in the market value of the asset. Bank B will pay the LIBOR-based return; it will also pay any difference if there is a depreciation in the price of the asset. The economic effect is as if Bank B owned the underlying asset, as such TR swaps are synthetic loans or securities. A significant feature is that Bank A will usually hold the underlying asset on its balance sheet, so that if this asset was originally on Bank B's balance sheet, this is a means by which the latter can have the asset removed from its balance sheet for the term of the TR swap.⁵ If we assume Bank A has access to LIBOR funding, it will receive a spread on this from Bank B. Under the terms of the swap, Bank B will pay the difference between the initial market value and any depreciation, so it is sometimes termed the 'guarantor', while Bank A is the 'beneficiary'.

⁵ Although it is common for the receiver of the LIBOR-based payments to have the reference asset on its balance sheet, this is not always the case.

Figure I.B.6.5 Total return swap



The total return on the underlying asset is the interest payments and any change in the market value if there is capital appreciation. The value of an appreciation may be cash settled, or alternatively there may be physical delivery of the reference asset on maturity of the swap, in return for a payment of the initial asset value by the total return 'receiver'. The maturity of the TR swap need not be identical to that of the reference asset, and in fact it is rare for it to be so.

The swap element of the trade will usually pay on a quarterly or semi-annual basis, with the underlying asset being revalued or marked-to-market on the refixing dates. The asset price is usually obtained from an independent third party source such as Bloomberg or Reuters, or as the average of a range of market quotes. If the obligor of the reference asset defaults, the swap may be terminated immediately, with a net present value payment changing hands according to what this value is, or it may be continued with each party making appreciation or depreciation payments as appropriate. This second option is only available if there is a market for the asset, which is unlikely in the case of a bank loan. If the swap is terminated, each counterparty will be liable to the other for accrued interest plus any appreciation or depreciation of the asset. Commonly under the terms of the trade, the guarantor bank has the option to purchase the underlying asset from the beneficiary bank, then dealing directly with loan defaulter.

With a TRS the basic concept is that one party 'funds' an underlying asset and transfers the total return of the asset to another party, in return for a (usually) floating return that is a spread to LIBOR. This spread is a function of:

- the credit rating of the swap counterparty;
- the amount and value of the reference asset;

- the credit quality of the reference asset;
- the funding costs of the beneficiary bank;
- any required profit margin;
- the capital charge associated with the TR swap.

The TRS counterparties must therefore consider a number of risk factors associated with the transaction, which include:

- the probability that the TR beneficiary may default while the reference asset has declined in value;
- the reference asset obligor defaults, followed by default of the TR swap receiver before payment of the depreciation has been made to the payer or ‘provider’.

The first risk measure is a function of the probability of default by the TRS receiver and the market volatility of the reference asset, while the second risk is related to the joint probability of default of both factors as well as the recovery probability of the asset.

TRS contracts are used in a variety of applications by banks, other financial institutions and corporates. They can be written as pure exchanges of cash flow differences – rather like an interest-rate swap – or the reference asset can be actually transferred to the total return payer, which would then make the TRS akin to a ‘synthetic repo’ contract.⁶

- *As pure exchanges of cash flow differences* Using TRSs as a credit derivative instrument, a party can remove exposure to an asset without having to sell it. This is conceptually similar to interest-rate swaps, which enable banks and other financial institutions to trade interest-rate risk without borrowing or lending cash funds. A TRS agreement entered into as a credit derivative is a means by which banks can take on unfunded off-balance-sheet credit exposure. Higher-rated banks that have access to LIBID funding can benefit by funding on-balance-sheet assets that are credit protected through a credit derivative such as a TRS, assuming the net spread of asset income over credit protection premium is positive.
- *Reference asset transferred to the total return payer* In a vanilla TRS the total return payer retains rights to the reference asset, although in some cases servicing and voting rights may be transferred. The total return receiver gains an exposure to the reference asset without having to pay out the cash proceeds that would be required to purchase it. As the maturity of the

⁶ When a bank sells stock short, it must borrow the stock to deliver it to its customer, in return for a fee (called a stock loan), or it may lend cash against the stock which it then delivers to the customer (called a ‘sale and repurchase agreement’ or *repo*). The counterparty is ‘selling and buying back’ while the bank that is short the stock is ‘buying and selling back’. A TRS is a synthetic form of repo, as the bond is sold to the TRS payer.

swap rarely matches that of the asset, the swap receiver may gain from the positive funding or *carry* that derives from being able to roll over short-term funding of a longer-term asset.⁷ The total return payer, on the other hand, benefits from protection against market and credit risk for a specified period of time, without having to liquidate the asset itself. On maturity of the swap the total return payer may reinvest the asset if it continues to own it, or it may sell the asset in the open market. Thus the instrument may be considered a *synthetic repo*.

The economic effect of the two applications may be the same, but they are considered different instruments:

- The TRS as a credit derivative instrument actually takes the assets off the balance sheet, whereas the tax and accounting authorities treat repo as if the assets remain on the balance sheet.
- A TRS trade is conducted under the ISDA standard legal agreement, while repo is conducted under a standard legal agreement called the Global Master Repurchase Agreement (GMRA)

It is these differences that, under certain circumstances, make the TRS funding route a more favourable one.

We now explain in more detail the main uses of TRSs.

I.B.6.4.1 Synthetic Repo

A portfolio manager believes that a particular bond (which she does not hold) is about to decline in price. To reflect this view she may do one of the following.

- *Sell the bond in the market and cover the resulting short position in repo.* The cash flow out is the coupon on the bond, with capital gain if the bond falls in price. Assume that the repo rate is floating, say LIBOR plus a spread. The manager must be aware of the funding costs of the trade, so that unless the bond can be covered in repo at *general collateral rates*⁸, the funding will be at a loss. The yield on the bond must also be lower than the LIBOR plus spread received in the repo.
- *As an alternative, enter into a TRS.* The portfolio manager pays the total return on the bond and receives LIBOR plus a spread. If the bond yield exceeds the LIBOR spread, the funding will be negative; however, the trade will gain if the trader's view is proved correct and the bond

⁷ This assumes a positively sloping yield curve.

falls in price by a sufficient amount. If the breakeven funding cost (which the bond must exceed as it falls in value) is lower in the TRS, this method will be used rather than the repo approach. This is more likely if the bond is special.

I.B.6.4.2 Reduction in Credit Risk

A TRS conducted as a synthetic repo is usually undertaken to effect the temporary removal of assets from the balance sheet. This can be done by entering into a short-term TRS with say, a two-week term that straddles the reporting date. Bonds are removed from the balance sheet if they are part of a sale plus TRS transaction. This is because legally the bank selling the asset is not required to repurchase bonds from the swap counterparty, nor is the total return payer obliged to sell the bonds back to the counterparty (or indeed sell the bonds at all on maturity of the TRS).

Hence, under a TRS an asset such as a bond position may be removed from the balance sheet. This may be desired for a number of reasons, for example if the institution is due to be analysed by credit rating agencies or if the annual external audit is due shortly. Another reason why a bank may wish to temporarily remove lower credit-quality assets from its balance sheet is if it is in danger of breaching capital limits in between the quarterly return periods. In this case, as the return period approaches, lower-quality assets may be removed from the balance sheet by means of a TRS, which is set to mature after the return period has passed. In summary, to avoid adverse impact on regular internal and external capital and credit exposure reporting a bank may use TRSs to reduce the amount of lower-quality assets on the balance sheet.

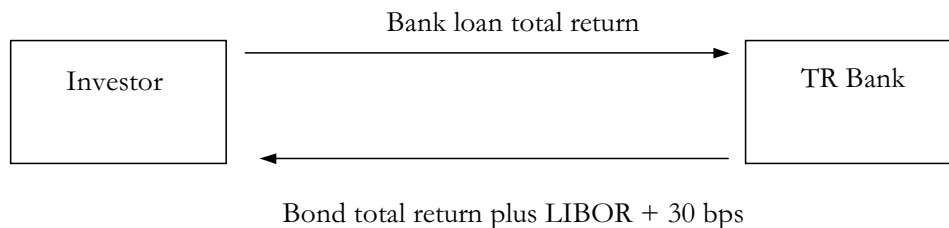
I.B.6.4.3 Capital Structure Arbitrage

A capital structure arbitrage describes an arrangement whereby investors exploit mispricing between the yields received on two different loans by the same issuer. Assume that the reference entity has both a commercial bank loan and a subordinated bond issue outstanding, but that the former pays LIBOR plus 330 basis points while the latter pays LIBOR plus 230 basis points. An investor enters into a TRS in which it is effectively purchasing the bank loan and selling short the bond. The nominal amounts will be at a ratio of, say, 2:1, as the bonds will be more price-sensitive to changes in credit status than the loans.

The trade is illustrated in Figure I.B.6.6. The investor receives the ‘total return’ on the bank loan, while simultaneously paying the return on the bond in addition to LIBOR plus 30 basis points, which is the price of the TRS. The swap generates a net spread of $(100 \text{ bps} \times \frac{1}{2}) + 250 \text{ bps} \times \frac{1}{2} = 175$ basis points.

⁸ That is, the bond cannot be *special*. A bond is special when the repo rate payable on it is significantly (say, 20-30 basis points or more) below the *general collateral* repo rate, so that covering a short position in the bond entails paying a

Figure I.B.6.6 Total return swap in capital structure arbitrage



I.B.6.4.4 The TRS as a Funding Instrument

A TRS can be regarded as a funding instrument, in other words as a substitute for a repo trade. There may be legal, administrative, operational or other reasons why a repo trade is not entered into to begin with. In these cases, provided that a counterparty can be found and the funding rate is not prohibitive, a TRS may be just as suitable.

Consider a financial institution such as a regulated broker-dealer that has a portfolio of assets on its balance sheet that it needs to obtain funding for. These assets are investment-grade structured finance bonds such as credit card asset-backed securities, residential mortgage-backed securities and collateralised debt obligation notes, and investment-grade convertible bonds. In the repo market, it is able to fund these at LIBOR plus 6 basis points. That is, it can repo the bonds out to a bank counterparty, and will pay LIBOR plus 6 bps on the funds it receives.

Assume that for operational reasons the bank can no longer fund these assets using repo. It can fund them using a basket TRS instead, providing a suitable counterparty can be found. Under this contract, the portfolio of assets is swapped out to the TRS counterparty, and cash received from the counterparty. The assets are therefore sold off the balance sheet to the counterparty, an investment bank. The investment bank will need to fund this itself, it may have a line of credit from a parent bank or it may swap the bonds out itself. The funding rate it charges the broker-dealer will depend on the rate at which it can fund the assets itself. Assume this is LIBOR plus 12 bps – the higher rate reflects the lower liquidity in the basket TRS market for non-vanilla bonds.

The broker-dealer enters into a three-month TRS with the investment bank counterparty, with a one-week interest-rate reset. This means that at each week interval the basket is revalued. The difference in value from the last valuation is paid (if higher) or received (if lower) by the investment bank to the broker-dealer; in return the broker-dealer also pays one-week interest on the funds it received at the start of the trade. In practice these two cash flows are netted off and

substantial funding premium.

only one payment changes hands, just as in an interest-rate swap. The terms of the trade are shown below.

- *Trade date* 22 December 2003
- *Value date* 24 December 2003
- *Maturity date* 24 March 2004
- *Rate reset* 31 December 2003
- *Interest rate* 1.19875% (*this is the one-week USD LIBOR fix of 1.07875 plus 12 bps*)

The swap is a three-month TRS with one-week reset, which means that the swap can be broken at one-week intervals and bonds in the reference basket can be returned, added to or substituted.

Assume that the portfolio basket contains five bonds, all US dollar denominated. Assume further that these are all investment-grade credit card asset-backed securities with prices available on Bloomberg. The combined market value of the entire portfolio is taken to be \$151,080,951.00.

At the start of the trade, the five bonds are swapped out to the investment bank, which pays the portfolio value for them. On the first reset date, the portfolio is revalued and the following calculations confirmed:

- *Old portfolio value* \$151,080,951.00
- *Interest rate* 1.19875%
- *Interest payable by broker-dealer* \$35,215.50
- *New portfolio value* \$152,156,228.00
- *Portfolio performance* + \$1,075,277
- *Net payment: broker-dealer receives* \$1,040,061.50

The rate is reset for value 31 December 2003 for the period to 7 January 2004. The rate is 12 bps over the one-week USD LIBOR fix on 29 December 2003, which is 1.15750 + 0.12 or 1.2775%. This interest rate is payable on the new 'loan' amount of \$152,156,228.00.

The TRS trade has become a means by which the broker-dealer can obtain collateralised funding for its portfolio. Like a repo, the bonds are taken off the broker-dealer's balance sheet, but unlike a repo the tax and accounting treatment also assumes they have been permanently taken off the balance sheet. In addition, the TRS is traded under the ISDA legal definitions, compared to a repo which is traded under the GMRA standard repo legal agreement.

I.B.6.5 Credit Options

Credit options are also bilateral over-the-counter financial contracts. A credit option is a contract designed to meet specific hedging or speculative requirements of an entity, which may purchase or sell the option to meet its objectives. A credit call option gives the buyer the right – without the obligation – to purchase the underlying credit-sensitive asset, or a credit spread, at a specified price and specified time (or period of time). A credit put option gives the buyer the right – without the obligation – to sell the underlying credit-sensitive asset or credit spread. By purchasing credit options banks and other institutions can take a view on credit spread movements for the cost of the option premium only, without recourse to actual loans issued by an obligor. The writer of credit options seeks to earn premium income.

Credit option terms are similar to those used for conventional equity options. A *call* option written on a stock grants the purchaser the right but not the obligation to purchase a specified amount of the stock at a set price and time. A credit option can be used by bond investors to hedge against a decline in the price of specified bonds, in the event of a credit event such as a ratings downgrade. The investor would purchase an option whose payoff profile is a function of the credit quality of the bond, so that a loss on the bond position is offset by the payout from the option.

As with conventional options, there are both vanilla credit options and exotic credit options. The vanilla credit option grants the purchaser the right, but not the obligation, to buy (or sell if a *put* option) an asset or credit spread at a specified price (the *strike* price) for a specified period of time up to the maturity of the option. A credit option allows a market participant to take a view on credit only, and no other exposure such as interest rates. As an example, consider an investor who believes that a particular credit spread, which can be that of a specific entity or the average for a sector (such as ‘all AA-rated sterling corporates’), will widen over the next six months. She can buy a six-month call option on the relevant credit spread, for which a one-off premium (the price of the option) is paid. If the credit spread indeed does widen beyond the strike during the six months, the option will be in the money and the investor will gain. If not, the investor’s loss is limited to the premium paid. Depending on whether the option is American or European, the option may be exercised before its expiry date or on its expiry date only.

Exotic credit options are options that have one or more of their parameters changed from the vanilla norm; the same terms are used as in other option markets. Examples include the barrier credit option, which specifies a credit event that would trigger (activate) the option or inactivate it. A digital credit option would have a payout profile that would be fixed, irrespective of how much in the money it was on expiry, and a zero payout if out of the money.

I.B.6.6 Synthetic Collateralised Debt Obligations

Credit derivatives are the key ingredient in the composition of structured credit products. These are typified by synthetic collateralised debt obligations (CDOs). Synthetic CDOs differ from conventional ‘cash’ CDOs in that they do not involve an actual transfer of underlying assets. In this section we discuss synthetic CDOs; to fully understand them, though, it is necessary first to describe cash CDOs.

I.B.6.6.1 Cash Flow CDOs

A cash flow CDO is not a credit derivative. It is a structured finance product in which a distinct legal entity known as a special-purpose vehicle issues bonds or notes against an investment in cash flows of an underlying pool of assets. These assets can be bonds, commercial bank loans or a mixture of both bonds and loans. Originally CDOs were developed as repackaging structures for high-yield bonds and illiquid instruments such as certain convertible bonds. Banks and financial institutions use CDOs to diversify their sources of funding, to manage portfolio risk and to obtain regulatory capital relief.

There are two types of CDO: *collateralised bond obligations* (CBOs) and *collateralised loan obligations* (CLOs). As the names suggest, the primary difference between each type is the nature of the underlying assets: a CBO will be collateralised by a portfolio of bonds while a CLO will represent an underlying pool of bank loans. The mechanics involved in structuring a CDO are similar in many respects to more traditional asset-backed securities. Das (2001) identifies the following areas of commonality:

- The originator of the transaction (called the *originator* or *sponsor* of the CDO)⁹ establishes a bankruptcy-remote legal entity known as a special-purpose vehicle that is the formal issuer of the notes.¹⁰
- It is the SPV that formally purchases the assets and their associated cash flows from the originator, thus taking the assets off the latter’s balance sheet; this is viewed as a ‘true sale’, that is, in legal and practical terms a completely separate legal entity (the SPV) has actually purchased the assets from the sponsor.
- The funds used to purchase the assets are raised through the issue of bonds into the debt capital market, which may be in more than one tranche and include an equity piece that is usually retained by the sponsor.

⁹ In this chapter we use the term ‘sponsor’ not ‘originator’ – but both terms are in common use.

¹⁰ As the assets of the SPV, which issues the notes, are actually sold to it from the sponsor, they become ring-fenced from it and are not affected by any changes in the sponsor’s fortunes. As such they would not be affected if the sponsor were to file for bankruptcy, hence ‘bankruptcy-remote’.

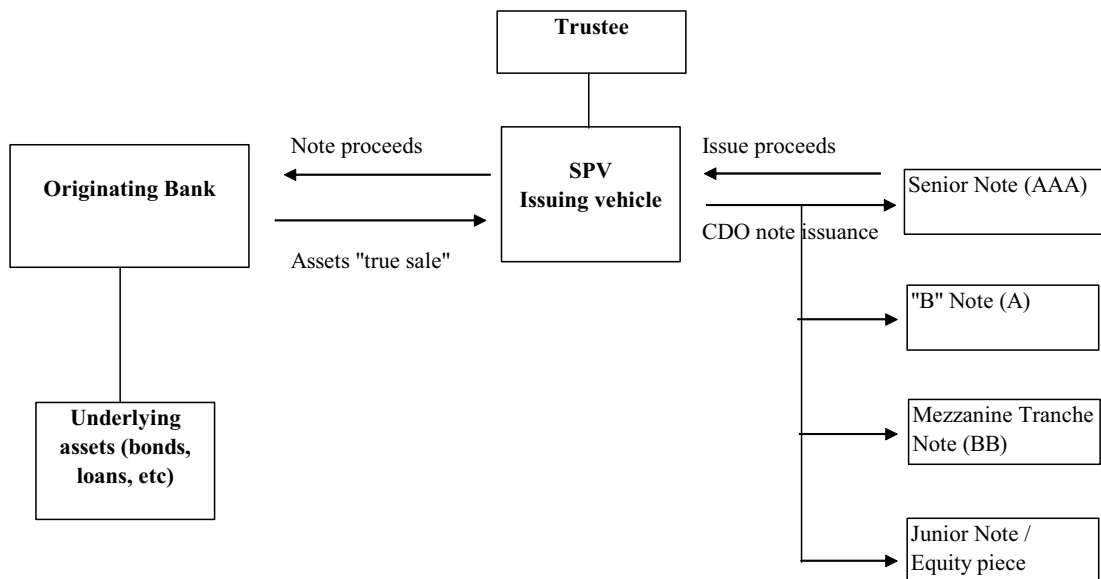
- Tranched securities are generally rated by a rating agency, with the rating reflecting the credit quality of the underlying assets as well as any measures put in place to reduce credit risk, known as *credit enhancement*.
- Investors purchasing the issued notes can expect to receive interest and principal payments as long as the underlying asset pool does not experience default to any significant extent.

A typical cash flow CDO is similar to other asset-backed securitisations involving an SPV. Bonds or loans are pooled together and the cash flows from these assets used to back the liabilities of the notes issued into the market. As the underlying assets are sold to the SPV, they are removed from the sponsor's balance sheet; hence the credit risk associated with these assets is transferred to the holders of the issued notes. The sponsor also obtains funding by issuing the notes. The generic structure is illustrated at Figure I.B.6.7.

The credit risk exposure of the underlying pool of assets is packaged into varying levels of risk. The most risky note issued by the SPV is the lowest-rated one, or the non-rated 'equity' note. A typical note tranching of the pool of assets might be:

- senior note, AAA-rated, comprising 90–95% of total principal
- subordinated note, A-rated, comprising 3–5% of total principal
- mezzanine note, BBB-rated, comprising 1–3% of total principal
- equity note, non-rated, comprising 1–2% of total principal.

Figure I.B.6.7 Generic cash flow CDO



The cash flows of the underlying assets are used to fund the liabilities of the overlying notes. As the notes carry different ratings, there is a priority of payment that must be followed: the most senior payment must be paid in full before the next payment can be met, and so on until the most junior liability is discharged. If there are insufficient funds available, the most senior notes must be paid off before the junior liabilities can be addressed. Different risk profiles of the issued notes arise from this subordinated structure. In addition, the structure makes use of *credit enhancements* to varying degrees, which include the following:

- *Overcollateralisation* The overlying notes are lower in value compared to the underlying pool; for example, \$250m nominal of assets are used as backing for \$170m nominal of issued bonds.
- *Cash reserve accounts* A reserve is maintained in a cash account and used to cover initial losses; the funds may be sourced from part of the proceeds.
- *Excess spread* Cash inflows from assets that exceed the interest service requirements of liabilities.
- *Insurance wraps* Insurance cover against losses suffered by the asset pool, for which an insurance premium is paid for as long as the cover is needed.

In addition to these credit enhancements, the quality of the collateral pool is monitored regularly and reported on by the portfolio administrator, who produces an investor report. This report details the results of various compliance tests, which are undertaken at individual asset level as well as aggregate level, and confirm the continuing quality of the asset pool.

I.B.6.6.2 What is a Synthetic CDO?

A synthetic CDO is so called because the transfer of credit risk is achieved ‘synthetically’. Conventional cash flow deals feature an actual transfer of ownership of the underlying loans or bonds to a separately incorporated legal entity, the SPV. In a synthetic CDO a *synthetic securitisation* structure is engineered so that only the credit risk of the assets is transferred by the sponsor, from itself to the investors, by means of credit derivative instruments. The loans or bonds themselves are not legally transferred. They normally remain on the sponsor’s balance sheet. It is just the credit risk of the asset pool (now known as the *reference portfolio*) that is transferred, using credit derivatives.

The sponsor is the credit protection buyer and the investors are the credit protection sellers.

A synthetic CDO is often constructed out of the following:

- a short position in a credit default swap (bought protection), by which the sponsor transfers its portfolio credit risk to the investor by means of the credit derivative;

- a long position in a portfolio of bonds or loans, the cash flow from which enables the sponsor to pay interest on overlying notes.

They are thus a natural progression of credit derivative structures, with single-name credit default swaps being replaced by *portfolio default swaps*.

Typically a large majority of the credit risk is transferred via notes that are sold to a wide set of investors and the proceeds are invested in risk-free collateral such as Treasury bonds. The most junior note, known as the *first-loss piece*, is often retained by the sponsor. Thus, on occurrence of a credit event among the reference assets, the originating bank receives funds remaining from the collateral only after they have been used to pay the principal on the issued notes, less the value of the junior note.

I.B.6.6.3 Funding Synthetic CDOs

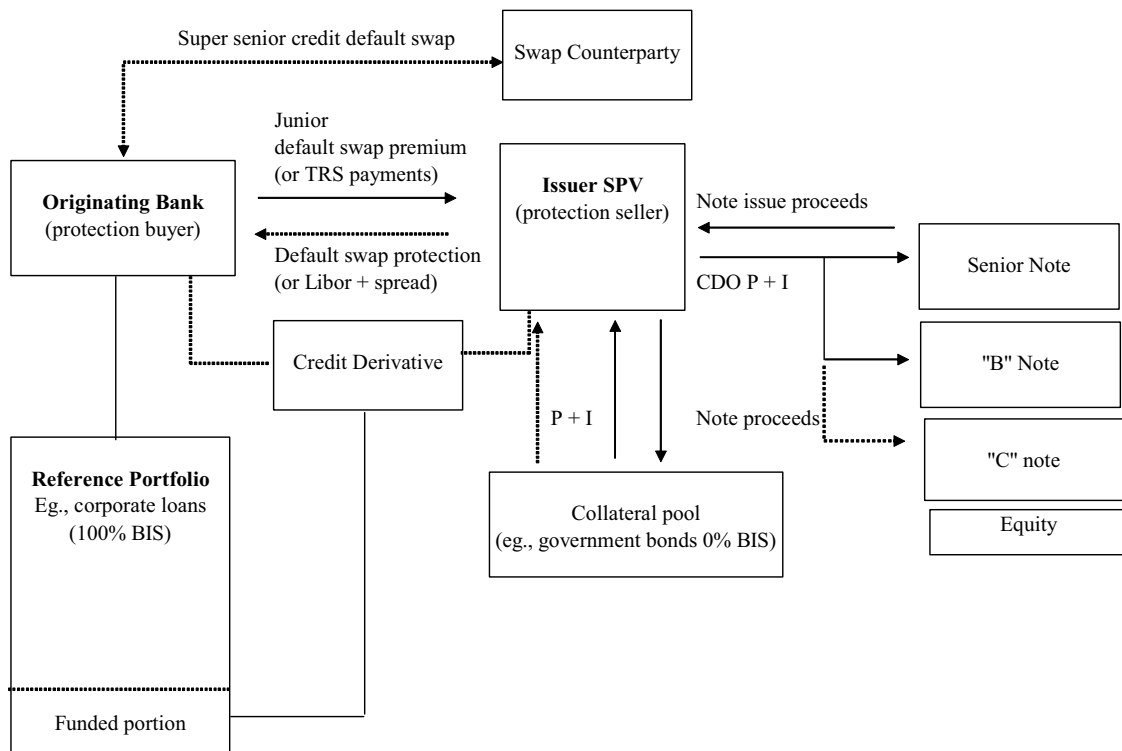
Synthetic CDOs can be unfunded, partially funded or fully funded. In their partially funded form they can be ‘de-linked’ from the sponsoring institution so that investors do not have any credit exposure to the sponsor.

- An *unfunded synthetic CDO* is a completely unfunded structure which uses CDSs to transfer the entire credit risk of the reference assets to investors who are protection sellers.
- In a *partially funded synthetic CDO*, only the highest credit risk segment of the portfolio is transferred. The cash flow that would be needed to service the synthetic CDO overlying liability is received from the AAA-rated collateral that is purchased by the SPV with the proceeds of the note issue. The sponsor obtains maximum regulatory capital relief by means of a partially funded structure, through a combination of the synthetic CDO and what is known as a *super senior swap* arrangement with an OECD banking counterparty. A super senior swap provides additional protection to that part of the portfolio (i.e. the senior segment) that is already protected by the funded portion of the transaction. The sponsor may retain the super senior element or may sell it on to a CDS provider (or simply a monoline insurance firm).¹¹
- A *fully funded synthetic CDO* is a structure where the credit risk of the entire portfolio is transferred to the SPV via a CDS. In a fully funded (or just ‘funded’) synthetic CDO the issuer enters into the CDS with the SPV, which itself issues notes to the value of the assets on which the risk has been transferred. The proceeds from the notes are invested in risk-free government or agency debt or in senior unsecured bank debt. Should there be a default on one or more of the underlying assets, the required amount of the collateral is sold and the

proceeds from the sale paid to the issuer to recompense for the losses. The premium paid on the CDS must be sufficiently high to ensure that it covers the difference in yield between that on the collateral and that on the notes issued by the SPV.

- The *fully unfunded synthetic* CDO uses only credit derivatives in its structure. The swaps are rated in a similar fashion to notes, and there is usually an ‘equity’ piece that is retained by the sponsor. The reference portfolio will again be commercial loans, usually 100% risk-weighted, or other assets. The credit rating of the swap tranches is based on the rating of the reference assets, as well as other factors such as the diversity of the assets and ratings performance correlation. As well as the equity tranche, there will be one or more junior tranches, one or more senior tranches and a super senior tranche. The senior tranches are sold on to AAA-rated banks as a portfolio CDS, while the junior tranche is usually sold to an OECD bank. The ratings of the tranches will typically be: super senior, AAA; senior, AA to AAA; junior, BB to A; and equity, unrated.

Figure I.B.6.8 Partially funded synthetic CDO structure



¹¹ A number of investment-type companies provide guarantees on portfolios in return for a fee. Such firms are known as monoline insurance companies, and are frequently the counterparties to the super senior swap element of a synthetic CDO.

A generic synthetic CDO structure is shown in Figure I.B.6.8. The credit risk of the reference assets is transferred to the issuer SPV and ultimately the investors, by means of the CDS and an issue of CLNs. In the default swap arrangement, the risk transfer is undertaken in return for the swap premium, which is then paid to investors by the issuer. The note issue is invested in risk-free collateral rather than passed on to the sponsor, in order to de-link the credit ratings of the notes from the rating of the sponsor. If the collateral pool was not established, a downgrade of the sponsor could result in a downgrade of the issued notes. Investors in the notes expose themselves to the credit risk of the reference assets, and if there are no credit events they will earn returns at least the equal of the collateral assets and the default swap premium. If the notes are credit-linked, they will also earn excess returns based on the performance of the reference portfolio. If there are credit events, the issuer will deliver the assets to the swap counterparty and will pay the nominal value of the assets to the sponsor out of the collateral pool. Credit default swaps are unfunded credit derivatives, while CLNs are funded credit derivatives where the protection sellers (the investors) fund the value of the reference assets upfront, and will receive a reduced return on occurrence of a credit event.

I.B.6.6.4 Variations in Synthetic CDOs

Synthetic CDOs have been issued in a variety of forms, based on both arbitrage CDOs and balance-sheet CDOs. Most structures have a reasonable amount in common with each other, differing only in detail.

A *synthetic arbitrage* CDO is originated generally by collateral managers who wish to exploit the difference in yield between that obtained on the underlying assets and that payable on the CDO, both in note interest and servicing fees. The generic structure is as follows: the SPV enters into a total return swap with the originating bank or financial institution, referencing the bank's underlying portfolio (the reference portfolio).¹² The portfolio is actively managed and is funded on the balance sheet by the originating bank. The SPV receives the 'total return' from the reference portfolio, and in return it pays LIBOR plus a spread to the originating bank. The SPV also issues notes that are sold into the market to CDO investors, and these notes can be rated as high as AAA as they are backed by high-quality collateral, which is purchased using the note proceeds.

A *balance-sheet synthetic* CDO is employed by banks that wish to manage regulatory capital. As before, the underlying assets are bonds, loans and credit facilities originated by the issuing bank. In a balance-sheet CDO the SPV enters into a CDS agreement with the sponsor, again with the

specific collateral pool designated as the reference portfolio.¹³ The SPV receives the premium payable on the default swap, and thereby provides credit protection on the reference portfolio.

I.B.6.6.5 Use of Synthetic CDOs

Using a synthetic CDO, the sponsor can obtain regulatory capital relief (because reference assets that are protected by credit derivative contracts, and which remain on the balance sheet, will, under Basel rules, attract a lower regulatory capital charge) and manage the credit risk on its balance sheet, but will not be receiving any funding. In other words, a synthetic CDO structure enables sponsors to separate credit risk exposure and asset funding issues.

Synthetic CDOs were introduced to meet the needs of sponsors for whom credit risk transfer is more important than funding considerations. Thus they are often preferred to cash flow CDOs for risk management and regulatory capital relief purposes. For banking institutions they also enable loan risk to be transferred without selling the loans themselves, thereby allowing customer relationships to remain unaffected. Indeed, the sponsors of the first synthetic deals were banks who wished to manage the credit risk exposure of their loan books, without having to resort to the administrative burden of true sale cash securitisation.

The first deals were introduced (in 1998) at a time when widening credit spreads and the worsening of credit quality among originating firms meant that investors were sellers of cash CDOs that had retained a credit linkage to the sponsor. A synthetic arrangement also meant that the credit risk of assets that were otherwise not suited to conventional securitisation could be transferred, while assets (such as bank guarantees, letters of credit or cash loans that had some legal or other restriction on being securitised) were still retained on the balance sheet. Indeed, for this reason synthetic deals are more appropriate for assets that are described under multiple legal jurisdictions.

The economic advantage of issuing a synthetic versus a cash CDO can be significant. Put simply, the net benefit to the sponsor is the gain in regulatory capital cost minus the cost of paying for credit protection on the credit default swap side. In a partially funded structure, a sponsoring bank will obtain full capital relief when note proceeds are invested in 0% risk-weighted collateral

¹² When a TRS is used in a synthetic CDO it is a funded credit derivative because the market price of the reference asset is paid upfront by the SPV. Therefore a liquidity facility (provided by the sponsor) is needed by the SPV, which it will draw on whenever it purchases a TRS.

¹³ Credit default swaps are not single-name swaps, but are written on a class of debt. The advantage for the originator is that it can name the reference asset class to investors without having to disclose the name of specific loans. Default swaps are usually cash settled and not physically settled, so that the reference assets can be replaced with other assets if desired by the sponsor.

such as Treasuries or gilts. The super senior swap portion will carry a 20% risk weighting (as long as the counterparty is an OECD bank, which is invariably the case).

In fact, a moment's thought should make clear to us that a synthetic CDO should be cheaper than an equivalent cash CDO. When credit default swaps are used the sponsor pays only a basis point fee, which for an AAA security this might be in the range 10–30 bps, depending on the stage of the credit cycle. In a cash structure where bonds are issued, the cost to the sponsor will be the benchmark yield plus the credit spread, and this will be considerably higher compared to the default swap premium.

I.B.6.6.6 Advantages and Limitations of Synthetic Structures

The introduction of synthetic securitisation vehicles was in response to specific demands of sponsoring institutions, and presents certain advantages over traditional cash flow structures. The fundamental advantages include the following:

- Speed of implementation: a synthetic transaction can, in theory, be placed in the market sooner than a cash deal, and the time from inception to closure can be as low as four weeks, with average execution time of six to eight weeks compared to three or four months for the equivalent cash deal; this reflects the shorter period required to deal in CDSs and the fact that cash assets do not have to be sourced in the market first.
- There is no requirement to fund the super senior element.
- For many reference names the CDS is frequently cheaper than the same-name underlying cash bond.
- Transaction costs such as legal fees can be lower as there is no necessity to set up an SPV.
- Banking relationships can be maintained with clients whose loans need not be actually sold off the sponsoring entity's balance sheet.
- The range of reference assets that can be covered is wider, and includes undrawn lines of credit, bank guarantees and derivative instruments that would give rise to legal and true sale issues in a cash transaction.
- The use of credit derivatives introduces greater flexibility to provide tailor-made solutions for credit risk requirements.
- The cost of buying protection is usually lower as there is little or no funding element and the credit protection price is below the equivalent-rate note liability.

This does not mean that the cash transaction is now an endangered species. It retains certain advantages over synthetic deals, which include:

- no requirement for an OECD bank (the 20% BIS risk-weighted entity) to act as the swap counterparty to meet capital relief requirements;

- lesser capital relief available by choosing a counterparty with a higher risk weighting such as a non- OECD bank;
- larger potential investor base, as the number of counterparties is potentially greater (certain financial and investing institutions have limitations on the usage of credit derivatives);
- lesser degree of counterparty exposure for originating entity. In a synthetic deal the default of a swap counterparty would mean cessation of premium payments or, more critically, a credit event protection payment, and termination of the CDS.

Investment banking advisers will structure the arrangement for their sponsoring client that best meets the latter's requirements. Depending on their nature, either a synthetic or a cash deal may be chosen.

I.B.6.7 General Applications of Credit Derivatives

Credit derivatives have allowed market participants to separate and disaggregate credit risk, and thence to trade this risk in a secondary market (see, for example, Das, 2000). Initially portfolio managers used them to reduce credit exposure; subsequently they have been used in the management of portfolios, to enhance portfolio yields and in the structuring of synthetic CDOs. Banks use credit derivatives to transfer credit risk of their loan and other asset portfolios, and to take on credit exposure based on their views on the credit market. In this regard they also act as credit derivatives market makers, running mismatched books in long- and short-position CDSs and TRSs. This is exactly how they operate in the interest-rate market, using interest-rate swaps.

I.B.6.7.1 Use of Credit Derivatives by Portfolio Managers

I.B.6.7.1.1 Enhancing portfolio returns

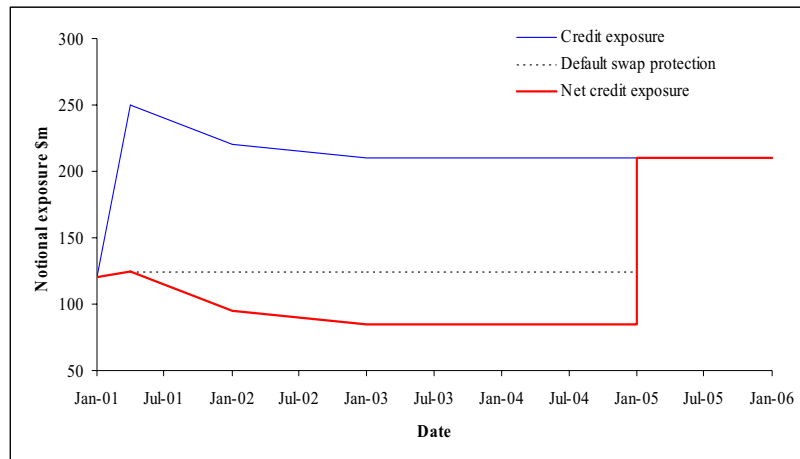
Asset managers can derive premium income by trading credit exposures in the form of derivatives issued with synthetic structured notes. This would be part of a structured credit product. A pool of risky assets can be split into specific tranches of risk, with the most risky portion given the lowest credit rating in the structure. This is known as 'multi-tranching'. The multi-tranching aspect of structured products enables specific credit exposures (credit spreads and outright default), and their expectations, to be sold to meet specific areas of demand. By using structured notes such as CLNs, tied to the assets in the reference pool of the portfolio manager, the trading of credit exposures is crystallised as added yield on the asset manager's fixed-income portfolio. In this way the portfolio manager enables other market participants to gain an exposure to the credit risk of a pool of assets but not to any other aspects of the portfolio, and without the need to hold the assets themselves.

I.B.6.7.1.2 Reducing credit exposure

Consider a portfolio manager who holds a large portfolio of bonds issued by a particular sector (say, utilities) and believes that spreads in this sector will widen in the short term. Previously, in order to reduce her credit exposure she would have to sell bonds; however, this may crystallise a mark-to-market loss and may conflict with her long-term investment strategy. An alternative approach would be to enter into a CDS, purchasing protection for the short term; if spreads do widen these swaps will increase in value and may be sold at a profit in the secondary market. Alternatively the portfolio manager may enter into total return swaps on the desired credits. She pays the counterparty the total return on the reference assets, in return for LIBOR. This transfers the credit exposure of the bonds to the counterparty for the term of the swap, in return for the credit exposure of the counterparty.

Consider now the case of a portfolio manager wishing to mitigate credit risk from a growing portfolio (say, one that has just been launched). Figure I.B.6.9 shows an example of an unhedged credit exposure to a hypothetical credit-risky portfolio. It illustrates the manager’s expectation of credit risk building up to \$250m as assets are purchased, and then reducing to a more stable level as the credits become more established.¹⁴ A three-year CDS entered into shortly after provides protection on half of the notional exposure, shown as the broken line. The net exposure to credit events has been reduced by a significant margin.

Figure I.B.6.9 Reducing credit exposure



I.B.6.7.1.3 Credit switches and zero-cost credit exposure

Protection buyers utilising CDSs must pay premium in return for laying off their credit risk exposure. An alternative approach for an asset manager involves the use of credit switches for

¹⁴ For instance, the fund may be invested in new companies. As the names become more familiar to the market the credits become more ‘established’ because the perception of how much credit risk they represent falls.

specific sectors of the portfolio. In a credit switch the portfolio manager purchases credit protection on one reference asset or pool of assets, and simultaneously sells protection on another asset or pool of assets.¹⁵ So, for example, the portfolio manager would purchase protection for a particular fund and sell protection on another. Typically the entire transaction would be undertaken with one investment bank, which would price the structure so that the net cash flows would be zero. This has the effect of synthetically diversifying the credit exposure of the portfolio manager, enabling her to gain and/or reduce exposure to sectors as desired.

I.B.6.7.1.4 Exposure to market sectors

Investors can use credit derivatives to gain exposure to sectors for which they do not wish a cash market exposure. This can be achieved with an *index* swap, which is similar to a TRS, with one counterparty paying a total return that is linked to an external reference index. The other party pays a LIBOR-linked coupon or the total return of another index. Indices that are used might include the government bond index, a high-yield index or a technology stocks index. Assume that an investor believes that the bank loan market will outperform the mortgage-backed bond sector; to reflect this view he enters into an index swap in which he pays the total return of the mortgage index and receives the total return of the bank loan index.

Another possibility is synthetic exposure to foreign currency and money markets. Again we assume that an investor has a particular view on an emerging market currency. If he wishes he can purchase a short-term (say one-year) domestic coupon-bearing note, whose principal redemption is linked to a currency factor. This factor is based on the ratio of the spot value of the foreign currency on issue of the note to the spot value on maturity. Such currency-linked notes can also be structured so that they provide an exposure to sovereign credit risk. The downside of currency-linked notes is that if the exchange rate goes the other way, the note will have a zero return, in effect a negative return once the investor's funding costs have been taken into account.

I.B.6.7.1.5 Trading Credit spreads

Assume that an investor has negative views on a certain emerging-market government bond credit spread relative to UK gilts. The simplest way to reflect this view would be to go long a CDS on the sovereign, paying X basis points. Assuming that the investor's view is correct and the sovereign bonds decrease in price as their credit spread widens, the premium payable on the credit swap will increase. The investor's swap can then be sold into the market at this higher premium.

¹⁵ A pool of assets would be concentrated on one sector, such as utility company bonds.

I.B.6.7.2 Use of Credit Derivatives by Banks

Banks use credit derivatives in exactly the same manner as portfolio managers – that is, in all the above we can replace ‘fund managers’ or ‘investors’ with ‘banks’. But in fact banks were the *first* users of credit derivatives. The market developed as banks sought to protect themselves from loss due to default on portfolios of mainly illiquid assets, such as corporate loans and emerging-market syndicated loans. Whilst securitization was a well-used technique to move credit risk off the balance sheet, often this caused relationship problems with obligors, who would feel that their close relationship with their banker was being compromised if the loans were sold off the bank’s balance sheet. Banks would therefore buy protection on the loan book using CDSs, enabling them to hedge their credit exposure whilst maintaining banking relationships. The loan would be maintained on the balance sheet but would be fully protected by the CDSs.

To illustrate, consider Figure I.B.6.10 which is a Bloomberg description page for a loan in the name of Haarman & Reimer, a chemicals company rated A3 by Moodys. We see that this loan pays 225 bps over LIBOR. Figure I.B.6.11 shows the CDS prices page for A3-rated Chemicals entities: Akzo Nobel is trading at 28 bps (to buy protection) as at 9 March 2004. A bank holding this loan can protect against default by purchasing this credit protection, and the relationship manager does not need to divulge this to the obligor. (In fact we may check the current price of this loan in the secondary market on the page BOAL, the Bank of America loan trading page on Bloomberg.)

Figure I.B.6.10 Haarman & Reimer loan description

Issue Information		Bank Group	Info @ Close
Borrower	HAARMANN & REIMER	Ld Arranger COBA,JPM	EURIBOR +225.000BP
Industry	Chemicals - Diversified	Agent	
Calc Type	(99) *NO CALCULATIONS*	Participants	55<GO>
Fac/Trnch Amts	EUR 880MM /400MM	Assignment Info	
Purpose	LBO	Min Pc	
Signing Date	11/28/02	Increment	
Effective Date	08/22/02	Fee	
Outstanding	400MM	Retain	Current Sprd & Fees
		Tranche Ratings	Interest Typ FLOATER
		S&P	NR
		Moody's	NR
		FI	NR
		Senior Debt Ratings	Current Base EURIBOR
Sub Limit Borrowings	Not Applicable	S & P	A+
		MOODY	A3
SR RTGS REFLECT: BAYER AG. TOTAL FAC INCLUDES AN ADDL €240MM MEZZANINE LOAN			

© Bloomberg LP. Used with permission.

Figure I.B.6.11 Chemicals sector CDS prices for Banco Bilbao Vizcaya, 9 March 2004

CHEMICALS/ PHARMACEUTICALS		3 Y - CDS Quotes			5 Y - CDS Quotes			TIME		
	BID	/	ASK	CHG	BID	/	ASK	CHG	TIME	
AKZO NOBEL	11	20	/	28	7:29	15	43	/	53	7:29
AVENTIS	2	13	/	23	7:29	16	24	/	34	7:29
BASF	3	10	/	17	7:29	17	10	/	20	7:29
BAYER	4	30	/	42	7:29	18	43	/	53	7:29
DEGUSSA	5	10	/	23	7:29	19	24	/	31	7:29
DSM	6	10	/	23	7:29	20	27	/	37	7:29
GSK	7	2	/	12	7:29	21	8	/	18	7:29
HENKEL KGAA	8	23	/	33	7:29	22	35	/	45	+2 12:28
ICI	9	55	/	75	7:29	23	80	/	90	7:29
LINDE	10	25	/	35	7:29	24	40	/	50	7:29
NOVARTIS	11	2	/	12	7:29	25	6	/	16	7:29
SOLVAY	12	/	/	/	7:29	26	25	/	32	7:29
SVENSKA AB	13	/	/	/	7:29	27	25	/	32	7:29
SYNGENTA AG	14	/	/	/	7:29	28	24	/	34	7:29

Tel: +34 91 537 6087
 INDICATIVE PRICES FOR CREDIT DEFAULT SWAPS ON STANDARD
 ISDA 2003 DOCUMENTATION WITH 3 CREDIT EVENTS
 MATURITIES ARE ON QUARTERLY BASIS

BBVA

Australia 61 2 9777 8600 Brazil 5511 3048 4500 Europe 44 20 7330 7500 Germany 49 69 920410
 Hong Kong 852 2977 6000 Japan 81 3 3201 8900 Singapore 65 6212 1000 U.S. 1 212 318 2000 Copyright 2004 Bloomberg L.P.
 6657-802-3 09-Mar-04 14:11:38

© Bloomberg L.P. © BBVA. Used with permission

The other major use by banks of credit derivatives is as a product offering for clients. The CDS market has developed exactly as the market did in interest-rate swaps, with banks offering two-way prices to customers and other banks as part of their product portfolio. Most commercial banks now offer this service, as they do in interest-rate swaps. In this role banks are both buyers and sellers of credit protection. Their net position will reflect their overall view on the market as well the other side of their customer business.

I.B.6.8 Unintended Risks in Credit Derivatives

As credit derivatives can be tailored to specific requirements in terms of reference exposure, term to maturity, currency and cash flows, they have enabled market participants to establish exposure to specific entities without the need for them to hold the bond or loan of that entity. This has raised issues of the different risk exposure that this entails compared to the cash equivalent. A recent Moody's special report (Tolk, 2001) highlights the unintended risks of holding credit exposures in the form of default swaps and credit-linked notes. Under certain circumstances it is possible for credit default swaps to create unintended risk exposure for holders, by exposing them to greater frequency and magnitude of losses compared to that suffered by a holder of the underlying reference credit.

In a credit default swap, the payout to a buyer of protection is determined by the occurrence of credit events. The definition of a credit event sets the level of credit risk exposure of the protection seller. A wide definition of 'credit event' results in a higher level of risk. To reduce the

likelihood of disputes, counterparties can adopt the ISDA definitions of credit derivatives to govern their dealings. The Moody's paper states that the current ISDA definitions do not unequivocally separate and isolate credit risk, and in certain circumstances credit derivatives can expose holders to additional risks. A reading of the paper would appear to suggest that differences in definitions can lead to unintended risks being taken on by protection sellers. Two examples from the paper are cited below as illustration.

Example I.B.6.2: Extending loan maturity

The bank debt of Consecoco, a corporate entity, was restructured in August 2000. The restructuring provisions included deferment of the loan maturity by three months, higher coupon, corporate guarantee and additional covenants. Under the Moody's definition, as lenders received compensation in return for an extension of the debt, the restructuring was not considered to be a 'diminished financial obligation', although Consecoco's credit rating was downgraded one notch. However, under the ISDA definition the extension of the loan maturity meant that the restructuring was considered to be a credit event, and thus triggered payments on default swaps written on Consecoco's bank debt. Hence, this was an example of a loss event under ISDA definitions that was not considered by Moody's to be a default.

Example I.B.6.3: Risks of synthetic positions and cash positions compared

Consider two investors in XYZ, one of whom owns bonds issued by XYZ while the other holds a credit-linked note referenced to XYZ. Following a deterioration in its debt situation, XYZ violates a number of covenants on its bank loans, but its bonds are unaffected. XYZ's bank accelerates the bank loan, but the bonds continue to trade at 85 cents on the dollar, coupons are paid and the bond is redeemed in full at maturity. However, the default swap underlying the CLN cites 'obligation acceleration' (of either bond or loan) as a credit event, so the holder of the CLN receives 85% of par in cash settlement and the CLN is terminated. However, the cash investor receives all the coupons and the par value of the bonds on maturity.

These two examples illustrate how, as CDSs are defined to pay out in the event of a very broad range of definitions of a 'credit event', portfolio managers may suffer losses as a result of occurrences that are not captured by one or more of the ratings agencies' rating of the reference asset. This results in a potentially greater risk for the portfolio manager compared to the position were it to actually hold the underlying reference asset. Essentially, therefore, it is important for the range of definitions of a 'credit event' to be fully understood by counterparties, so that holders of default swaps are not taking on greater risk than is intended.

I.B.6.9 Summary

This chapter has introduced credit derivatives, their form and structure, and explained why they are used. We conclude that credit derivatives:

- are instruments designed to transfer credit risk from one party to another;
- act as insurance contracts against loss suffered due to ‘credit events’; and
- exist in funded and unfunded forms, the former typified by credit linked notes and the latter by credit default swaps.

Credit events are specified in the contract legal documentation and are the trigger event upon which a payout is made under the credit derivative contract. They include bankruptcy, failure to pay, liquidation and actual default. Credit derivatives are used in a range of applications by banks, fund managers, non-bank financial institutions and corporations. These applications include:

- removing credit risk from bank balance sheet, without removing the risky assets themselves;
- entering into credit speculation by taking on credit risk exposure synthetically;
- acting as a credit derivatives market maker;
- gaining exposure to credit risky assets without having to acquire assets directly; and
- being used to structure structured credit products such as synthetic collateralised debt obligations.

As credit derivatives are over-the-counter products, they may be tailored to meet specific individual client requirements. This flexibility has been a significant factor in the rapid growth in their use.

References

Das S (2001) *Credit Derivatives and Credit Linked Notes*, 2nd edition (Singapore: Wiley), Chapters 2–4.

Francis J, J Frost and G Whittaker (1999), *Handbook of Credit Derivatives*, (New York: Irwin)

Tolk J (2001) ‘Understanding the Risks in Credit Default Swaps’, *Moody’s Investors Service Special Report*, 16 March.

Choudhry M (2004), *Structured Credit Products: Credit Derivatives and Synthetic Securitisation*, (Singapore: Wiley)

I.B.7 Caps, Floors and Swaptions

Lionel Martellini and Phillippe Priaulet¹

In this chapter we introduce caps, floors and swaptions. These are plain-vanilla (or standard) interest-rate options that are traded in over-the-counter contracts. Caps and floors are fixed-income securities designed to hedge interest-rate risk. The buyer of a cap is hedged against an increase in interest rates while the buyer of a floor is hedged against a decrease in interest rates. Swaptions allow the holder to enter some pre-specified underlying swap contract on or up to a pre-specified date, which is the expiration date of the swaption. Like caps and floors, swaptions are fixed-income securities designed to hedge interest-rate risk.

This chapter explains the structure of caps, floors and swaptions and how these instruments are used by practitioners to manage and speculate on interest-rate risk. We also examine the pricing of these instruments using the Black (1976) model and the manner in which prices are quoted in the market place.

I.B.7.1 Caps, Floors and Collars: Definition and Terminology

A *cap* is an over-the-counter contract by which the seller agrees to pay a positive amount to the buyer of the contract if the *reference rate* exceeds a pre-specified level called the *exercise rate* of the cap on given future dates. Conversely, the seller of a *floor* agrees to pay a positive amount to the buyer of the contract if the reference rate falls below the exercise rate on some future dates. We first define some terms:

- The *notional* or nominal amount is fixed in general.
- The *reference rate* is an interest-rate benchmark based for example on Libor, T-bill and T-bond yield to maturity, or swap rates, and from which the contractual payments are determined. The most usual ones are the one-month, three-month, six-month and one-year Libor rates, the constant maturity Treasury and constant maturity swap rates.
- The *exercise rate* or *strike rate* is a fixed rate determined at the origin of the contract.

¹ Lionel Martellini is a Professor of Finance at EDHEC Graduate School of Business, and the Scientific Director of EDHEC Risk and Asset Management Research Center. Phillippe Priaulet is a Fixed-Income Strategist, in charge of derivatives strategies for HSBC, and also an Associate Professor in the Department of Mathematics of the University of Evry Val d'Essonne.

*Reproduced with kind permission of John Wiley & Sons Ltd from *Fixed-Income Securities: Valuation, Risk Management and Portfolio Strategies*, 2003.

- The *settlement frequency* refers to the frequency with which the reference rate is compared to the exercise rate. The time between two payments is known as the *tenor*. It is expressed in years. The most common frequencies are monthly, quarterly, semi-annually and annually.
- The *starting date* is the date when the protection of caps, floors and collars begins.
- The *maturity* of caps, floors and collars can range from several months to 30 years, although liquidity exists generally only for maturities up to about 10 years or less, depending on the market.
- The *premium* of caps, floors and collars is the price paid. It is expressed as a percentage of the notional amount.

We give below two examples of caps and floors.

Example I.B.7.1

Let us consider a cap with a nominal amount of \$1,000,000, an exercise rate E , based upon a Libor rate with a δ -month maturity denoted $R^L(t, \delta)$ at date t , and with the following schedule of cash flows:

t	T_0	T_1	T_2	T_n
<i>Cash Flow</i>		C_1	C_2			C_n

T_0 is the starting date of the cap and the difference $T_n - T_0$ (expressed in years) is the maturity of the cap. For all $j = 1, \dots, n$, we assume a constant tenor $T_j - T_{j-1} = \delta$.² On each date T_j the cap holder receives a cash-flow C_j given by:

$$C_j = \$1,000,000 \times \delta \times [R^L(T_{j-1}, \delta) - E]^+$$

where $[x]^+ = \max(x, 0)$.

C_j is a call option on the Libor rate $R^L(t, \delta)$ observed on date T_{j-1} with a payoff occurring on date T_j . The cap is a portfolio of n such options. The n call options of the cap are known as the *caplets*.

Let us now consider a floor with the same characteristics. The floor holder gets on each date T_j (for $j = 1, \dots, n$) a cash flow F_j given by:

$$F_j = \$1,000,000 \times \delta \times [E - R^L(T_{j-1}, \delta)]^+$$

² Although actual day count is used in practice, here we simplify to avoid complex notation.

F_j is a put option on the Libor rate $R^L(t, \delta)$ observed on date T_{j-1} with a payoff occurring on date T_j . The floor is a portfolio of n such options. The n put options of the cap are known as the *floorlets*.

A *collar* is a combination of a cap and a floor. There are two kinds of collar: one involves buying a cap and selling a floor at the same time; the other involves buying a floor and selling a cap at the same time. In the former case the idea is that the premium of the floor reduces the cost of the cap; in the latter case the premium of the cap reduces the cost of the floor. We will see concrete examples of these two products in Section I.B.7.3.

I.B.7.2 Pricing Caps, Floors and Collars

The rapid rise of the markets for caps, floors and collars pre-dated the development of complex interest-rates models. In the absence of such models, the use of the Black (1976) model has become standard. This model, which is particularly tractable and simple to use, remains the reference for the market in terms of pricing and hedging standard assets such as caps, floors and swaptions, despite its simplifying assumptions. For example, the Black model, like the Black–Scholes–Merton formula upon which is based, assumes that the volatility of the underlying is constant. While this assumption is violated for all assets, it is particularly problematic for the volatility of the forward rate which becomes more sensitive to changes in the spot as maturity approaches.

I.B.7.2.1 Cap Formula

Let us consider a cap with nominal amount N and exercise rate E , based upon a reference linear rate denoted $R^L(t, \delta)$ at date t and with the schedule of cash flows shown in Example I.B.7.1. For all $j = 1, \dots, n$, we assume a constant tenor $T_j - T_{j-1} = \delta$. On each date T_j the cap holder receives a cash-flow C_j given by:

$$C_j = N \times \delta \times [R^L(T_{j-1}, \delta) - E]^+.$$

The cap price at date t in the Black (1976) model is given by an adaptation of the Black–Scholes–Merton formula for stock options (see Chapter I.A.8) as follows:

$$\begin{aligned} \text{Cap}(t) &= \sum_{j=1}^n \text{Caplet}_j(t) \\ &= \sum_{j=1}^n N \times \delta \times B(t, T_j) \times [F(t, T_{j-1}, T_j) \Phi(d_j) - E \Phi(d_j - \sigma_j \sqrt{T_{j-1} - t})] \end{aligned}$$

(I.B.7.1)

where

- Φ is the cumulative distribution function of the standard Gaussian distribution.
- $F(t, T_{j-1}, T_j)$ is the underlying of the caplet relative to the cash flow C_j received at date T_j . It is the forward linear rate as seen from date t , beginning at date T_{j-1} and finishing at date T_j . Note in particular that

$$F(T_{j-1}, T_{j-1}, T_j) = R^L(T_{j-1}, \delta)$$

- $B(t, T_j)$ is the discount rate from date T_j to date t (i.e., the price of a zero-coupon bond starting at date t and ending at date T_j). It can be calculated as:

$$B(t, T_j) = \frac{1}{1 + \delta F(t, T_{j-1}, T_j)}$$

- d_j is given by

$$d_j = \frac{\ln\left(\frac{F(t, T_{j-1}, T_j)}{E}\right) + 0.5\sigma_j^2(T_{j-1} - t)}{\sigma_j \sqrt{T_{j-1} - t}}$$

- σ_j is the volatility of the underlying rate $F(t, T_{j-1}, T_j)$ which is usually referred to as the caplet volatility.

Note that the first caplet yielding the cash flow C_1 is not taken into account in the cap price once the cap is initiated at date T_0 since, at that date, $R^L(T_0, \delta)$ is already known.

I.B.7.2.2 Floor Formula

Let us now consider a floor with nominal amount N and exercise rate E , based upon a reference linear rate denoted $R^L(t, \delta)$ at date t . For all $j = 1, \dots, n$, we assume a constant tenor $T_j - T_{j-1} = \delta$. On each date T_j the floor holder receives a cash flow F_j given by:

$$F_j = N \times \delta \times [E - R^L(T_{j-1}, \delta)]^+$$

The price at date t of a floor with the same parameters as those used for the cap is given by

$$\begin{aligned} \text{Floor}(t) &= \sum_{j=1}^n \text{Floorlet}_j(t) \\ &= \sum_{j=1}^n N \times \delta \times B(t, T_j) \times [-F(t, T_{j-1}, T_j)\Phi(-d_j) + E\Phi(-d_j + \sigma_j \sqrt{T_{j-1} - t})] \quad (\text{I.B.7.2}) \end{aligned}$$

The price at date t of a collar is simply the difference between the cap price and the floor price if we buy the cap and sell the floor. We obtain an analogous formula when we sell the cap and buy the floor. We refer the reader to Martellini *et al.* (2003) for more details on pricing and hedging of these instruments.

Example I.B.7.2

Price the following caplet which is designed to cap the interest on a loan of \$500,000 at 5.0% p.a., compounding quarterly. The caplet protects a three-month interest period starting six months from now. The forward interest rate for a three-month period starting on six months is 4.8% p.a., compounding quarterly, and the volatility of this rate is 18% p.a. So:

$$N = \$500,000$$

$$\delta = 0.25$$

$$F(t, T_{j-1}, T_j) = 0.048$$

$$E = 0.050$$

$$\sigma_j = 0.18$$

$$B(t, T_j) = \frac{1}{1 + (0.25 \times 0.048)} = 0.988142$$

$$d_j = \frac{\ln\left(\frac{0.048}{0.050}\right) + (0.5 \times 0.18^2 \times 0.5)}{0.18\sqrt{0.5}} = -0.257088$$

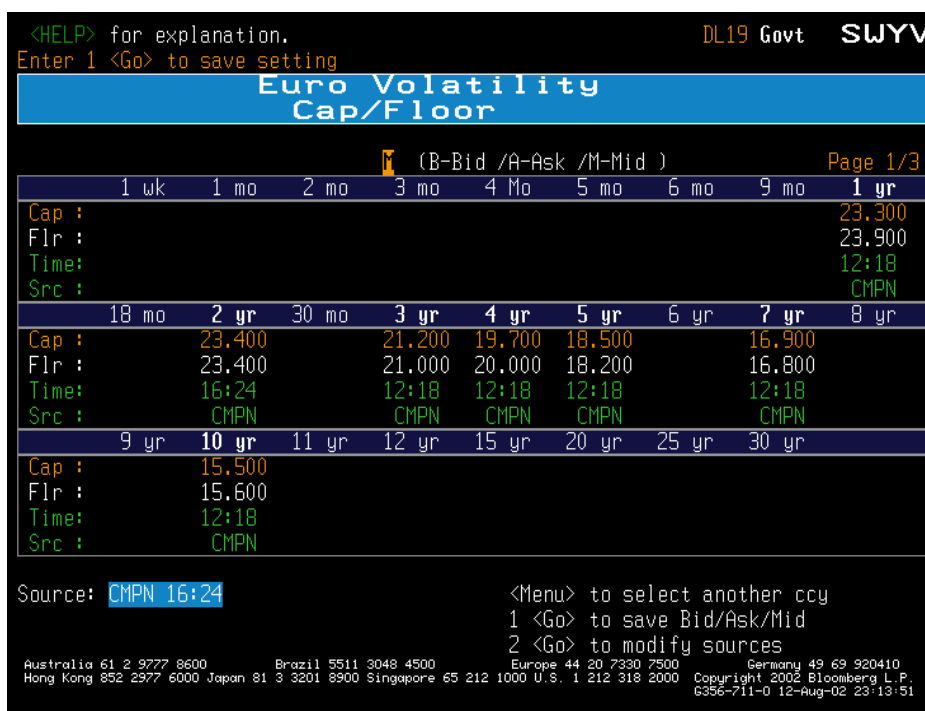
$$\begin{aligned} \text{Caplet} &= 500,000 \times 0.25 \times 0.988142 \times [0.048\Phi(-0.257088) - 0.050 \Phi(-0.257088 - 0.18\sqrt{0.5})] \\ &= \$199.23. \end{aligned}$$

I.B.7.2.3 Market Quotes

Cap and floor prices are usually quoted in terms of implied volatility. Given the market price of the cap or floor, one obtains the implied volatility by inversion of the Black (1976) formula (I.B.7.1) or (I.B.7.2). Conversely, given the implied volatility, one uses the formula (I.B.7.1) or (I.B.7.2) to obtain the market price.

The market reference is the implied volatility curve as a function of the maturity of the option. As an illustration, Figure I.B.7.1 shows the volatility curve for three-month Euribor caplets and floorlets, extracted from Bloomberg on 12 August 2002. The maturity of caps and floors ranges from 1 year to 10 years. The implied volatility quoted here is the mean of the bid and ask quotations. For example, the cap with five-year maturity has a 18.5% volatility and the floor with 10-year maturity has a 15.6% volatility.

Figure I.B.7.1: Volatility curve for three-month Euribor caps and floors



Source: Bloomberg Eurozone

I.B.7.3 Uses of Caps, Floors and Collars

Caps, floors and collars are fixed-income securities designed to hedge interest-rate risk. More precisely, a cap enables the buyer to cap the reference rate associated with a liability as a floor enables the buyer to protect the total return of assets. The buyer of a cap is hedged against an increase in interest rates (e.g., for hedging a floating-rate loan), while the buyer of a floor is hedged against a decrease in interest rates (e.g., for a floating-rate asset).

I.B.7.3.1 Limiting the Financial Cost of Floating-Rate Liabilities

Caps, and collars are options used to limit the financial cost of floating-rate liabilities. A collar combines a cap with a solid floor to reduce the premium.

Example I.B.7.3

On 12 April 2000, a firm which pays the six-month Libor semi-annually to a bank for the next two years with a \$1,000,000 notional amount wants to be hedged against a rise in this index. This firm decides to buy a cap with the following features:

- notional amount: \$1,000,000
- reference rate: six-month Libor
- strike rate: 5.5%
- starting date: 12 November 2000

- maturity: 2 years
- tenor: 6 months
- day-count: Actual/360

The premium from the buyer of the cap, which is paid on 6 November 2001, 12 November 2001, 6 November 2002 and 12 November 2002, is equal to 0.2% of the notional amount prorated to the period. Taking the point of view of the firm, the schedule for the cap is the following:

<i>Date j</i>	<i>Cash Flow</i>
12 th April 2000	Cap contract concluded
12 th November 2000	Starting date of cap
6 th November 2001	$10^6 \times (182/360) \times ([L_{j-1} - 5.5\%]^+ - 0.2\%)$
12 th November 2001	$10^6 \times (183/360) \times ([L_{j-1} - 5.5\%]^+ - 0.2\%)$
6 th November 2002	$10^6 \times (182/360) \times ([L_{j-1} - 5.5\%]^+ - 0.2\%)$
12 th November 2002	$10^6 \times (183/360) \times ([L_{j-1} - 5.5\%]^+ - 0.2\%)$

where L_j is the value of the six-month Libor on date j . Thus, for example, on 6 November 2002 the six-month Libor rate L_{j-1} is the rate on 12 November 2001 and on 12 November 2002 L_{j-1} is the rate on 6 November 2002.

If the six-month Libor is higher than 5.5%, the firm will receive the difference between the six-month Libor and 5.5%, and zero otherwise. By buying this cap, the firm has limited its financing cost to $5.5\% + 0.2\% = 5.7\%$ while benefiting from a lower cost of financing if the six-month Libor falls below 5.5%. In this case and if the six-month Libor remains stable at 3.5% during the life of the cap, the financing cost for the firm is equal to 3.7% (3.5% plus the premium).

The second example illustrates a collar (buy a cap and sell a floor) which limits the financing cost of floating-rate liabilities for a firm while reducing the cost of the hedge, i.e. the premium paid for the protection. But the drawback is that the firm will not benefit from a lower financing cost if the reference rate falls below the strike rate of the floor.

Example I.B.7.4

We consider the same assumptions as in the previous example. Now, however, the firm wants to reduce the hedge cost, and, for that purpose, sells at the same time a floor with the following features:

- notional amount: \$1,000,000

- reference rate: six-month Libor
- strike rate: 4.5%
- starting date: 12 November 2000
- maturity: 2 years
- tenor: 6 months
- day-count: Actual/360

The premium due by the buyer of the floor, which is paid on 6 November 2001, 12 November 2001, 6 November 2002 and 12 November 2002, is equal to 0.1% of the notional amount prorated to the period. Taking the point of view of the firm, the schedule for the collar is the following:

<i>Date j</i>	<i>Cash Flow</i>
12 April 2000	Cap and floor contracts are concluded
12 November 2000	Starting date of the cap and the floor
6 November 2001	$10^6 \times (182/360) \times ([L_{j-1} - 5.5\%]^+ - [4.5\% - L_{j-1}]^+ - 0.1\%)$
12 November 2001	$10^6 \times (183/360) \times ([L_{j-1} - 5.5\%]^+ - [4.5\% - L_{j-1}]^+ - 0.1\%)$
6 November 2002	$10^6 \times (182/360) \times ([L_{j-1} - 5.5\%]^+ - [4.5\% - L_{j-1}]^+ - 0.1\%)$
12 November 2002	$10^6 \times (183/360) \times ([L_{j-1} - 5.5\%]^+ - [4.5\% - L_{j-1}]^+ - 0.1\%)$

There are three different situations for the firm:

- If the six-month Libor is higher than 5.5%, the firm will receive the difference between the six-month Libor and 5.5%, thus offsetting its increased borrowing costs.
- If the six-month Libor is lower than 4.5%, the firm will pay the difference between 4.5% and the six-month Libor, thus eliminating the potential to benefit from lower rates.
- If the six-month Libor is between 4.5% and 5.5%, no cash flows are exchanged.

By contracting this collar, the firm has limited its hedge cost to $0.2\% - 0.1\% = 0.1\%$, 0.2% being the premium paid for buying the cap and 0.1% being the premium received for selling the floor. This collar enables the firm to limit its financing cost to $5.5\% + 0.1\% = 5.6\%$ while benefiting from a lower financing cost if the six-month Libor is between 4.5% and 5.5%. In this case and if the six-month Libor remains stable for example at 5% during the life of the collar, the financing cost for the firm is equal to 5.1% (5% plus the premium). But if the six-month Libor falls below 4.5%, the firm will not benefit from a lower financing cost. For example, if the six-month Libor remains stable at 3% during the life of the collar, the financing cost for the firm will be 4.6% (4.5% plus the premium).

I.B.7.3.2 Protecting the Rate of Return of a Floating-Rate Asset

Floors are options used to protect the rate of return of a floating-rate asset. Collars combine a floor with a short cap to reduce the cost of the hedge.

Example I.B.7.5

On 2 January 2001, a firm which has invested \$10,000,000 in the three-month Libor for the next year wants to be hedged against a decrease in this index. This firm decides to buy a floor with the following features:

- notional amount: \$10,000,000
- reference rate: three-month Libor
- strike rate: 4%
- starting date: 8 January 2001
- maturity: 1 year
- tenor: 3 months
- day-count: Actual/360

The premium due from the buyer of the floor, which is paid on 8 April 2001, 8 July 2001, 8 October 2001 and 8 January 2002, is equal to 0.2% of the notional amount prorated to the period. Taking the point of view of the firm, the schedule of the floor is the following:

<i>Date j</i>	<i>Cash Flow</i>
2 January 2001	Floor contract concluded
8 January 2001	Starting date of floor
8 April 2001	$10^7 \times (90/360) \times ([4\% - L_{j-1}]^+ - 0.2\%)$
8 July 2001	$10^7 \times (91/360) \times ([4\% - L_{j-1}]^+ - 0.2\%)$
8 October 2001	$10^7 \times (92/360) \times ([4\% - L_{j-1}]^+ - 0.2\%)$
8 January 2002	$10^7 \times (92/360) \times ([4\% - L_{j-1}]^+ - 0.2\%)$

where L_j is the value of the three-month Libor at date j .

If the three-month Libor is lower than 4%, the firm will receive the difference between 4% and the three-month Libor, and zero otherwise. By buying this floor, the firm has guaranteed a minimum rate of return of $4\% - 0.2\% = 3.8\%$ while benefiting from a better rate on his investment if the three-month Libor exceeds 4%. In this case, and if the three-month Libor remains stable at 6% during the life of the cap, the firm has guaranteed a rate of 5.8%.

The second example illustrates a collar (buy a floor and sell a cap) that protects the rate of return on a floating-rate asset while reducing the cost of the hedge, that is, the premium paid for the protection. The trade-off is that the firm will not benefit from better conditions on its investment if the reference rate goes up above the strike rate of the cap.

Example I.B.7.6

We consider the same assumptions as in the previous example. This time, however, the firm wants to reduce the cost of the hedge, and, for that purpose, sells at the same time a cap with the following features:

- notional amount: \$10,000,000
- reference rate: three-month Libor
- strike rate: 6.5%
- starting date: 8 January 2001
- maturity: 1 year
- tenor: 3 months
- day-count: Actual/360

The premium due from the buyer of the cap, which is paid on 8 April 2001, 8 July 2001, 8 October 2001 and 8 January 2002, is equal to 0.05% of the notional amount prorated to the period. Taking the point of view of the firm, the schedule of the collar is the following:

<i>Date j</i>	<i>Cash Flow</i>
2 January 2001	Cap and floor contracts are concluded
8 January 2001	Starting date of the cap and the floor
8 April 2001	$10^7 \times (90/360) \times ([4\% - L_{j-1}]^+ - [L_{j-1} - 6.5\%]^+ - 0.15\%)$
8 July 2001	$10^7 \times (91/360) \times ([4\% - L_{j-1}]^+ - [L_{j-1} - 6.5\%]^+ - 0.15\%)$
8 October 2001	$10^7 \times (92/360) \times ([4\% - L_{j-1}]^+ - [L_{j-1} - 6.5\%]^+ - 0.15\%)$
8 January 2002	$10^7 \times (92/360) \times ([4\% - L_{j-1}]^+ - [L_{j-1} - 6.5\%]^+ - 0.15\%)$

There are three different situations for the firm:

- If the three-month Libor is lower than 4%, the firm will receive the difference between 4% and the three-month Libor, thus compensating for lower investment returns.
- If the three-month Libor is higher than 6.5%, the firm will pay the difference between the three-month Libor and 6.5%, thus losing the benefits of higher returns.
- If the three-month Libor is between 4% and 6.5%, no cash flows are exchanged.

By contracting this collar, the firm has limited its hedging cost to $0.2\% - 0.05\% = 0.15\%$, 0.2% being the premium paid for buying the floor and 0.05% being the premium received for selling the cap. This collar enables the firm to guarantee a rate of return to $4\% - 0.15\% = 3.85\%$ while benefiting from a better rate of return if the three-month Libor is between 4% and 6.5% . In this case and if the three-month Libor remains stable for example at 5.5% during the life of the collar, the rate of return on the firm's investment is equal to 5.35% . But if the three-month Libor exceeds 6.5% , the firm will not benefit from a better rate of return. For example, if the three-month Libor remains stable for example at 8% during the life of the collar, the rate of return on the firm's investment will be 6.35% .

I.B.7.4 Swaptions: Definition and Terminology

Like caps and floors, swaptions are over-the-counter contracts. A European swaption is an option allowing the holder to enter some pre-specified underlying swap contract on a pre-specified date, which is the expiration date of the swaption. There are two kinds of European swaptions:

- The *receiver option on swap* is an option that gives the buyer the right to enter into a swap, receiving the fixed rate and paying the variable rate.
- The *payer option on swap* is an option that gives the buyer the right to enter into a swap, paying the fixed rate and receiving the variable rate.

We now introduce some terminology:

- The *exercise rate* or *strike rate* is the specified fixed rate at which the buyer can enter into the swap.
- The *maturity* or *expiry date* is the date when the option can be exercised. The maturity of swaptions can range from several months to 10 years.
- The *premium* of swaptions is the price, expressed as a percentage of the principal amount of the swap.
- *Bermudan* swaptions give to the buyer the opportunity to enter a swap on several specified dates in the future.
- *American* swaptions give to the buyer the opportunity to enter a swap at any time before the maturity date of the option.

Note that the underlying asset of the swaption is most commonly a plain vanilla swap whose maturity can range from 1 year to 30 years. We give below an example of a European swaption.

Example I.B.7.7: A Payer Swaption

We consider a standard underlying swap contract. The swaption with expiration date T_0 is defined by the following schedule:

<i>Fixed Leg</i>		$-F_1$	$-F_2$	$-F_n$
	T_0	T_1	T_2	T_n
<i>Variable Leg</i>		V_1	V_2	V_n

We assume that the buyer of the swaption has the right on date T_0 to enter a swap where he receives the variable leg and pays the fixed leg. For all $j = 1, \dots, n$, we have $T_j - T_{j-1} = \delta$, as before.

Cash flows for the fixed leg are on dates T_j for $j = 1, \dots, n$, and are given by:

$$-F_j = -\delta \times F \times N$$

where F is a fixed rate and N is the principal (i.e. the nominal amount in the swap).³ Cash flows for the floating leg are on dates T_j for $j = 1, \dots, n$, and are given by:

$$V_j = \delta \times R^L(T_{j-1}, \delta) \times N$$

where $R^L(T_{j-1}, \delta)$ is the Libor rate with a δ -month maturity at date T_{j-1} .

I.B.7.5 Pricing Swaptions

I.B.7.5.1 European Swaption Pricing Formula

We consider a standard payer swaption contract with the schedule shown above. Recall from Chapter I.B.4 that the value of a swap starting at date T_0 at any time $t < T_0$ is the present value of future cash flows. Thus, denoting by $B(t, T_j)$ the discount rate from date T_j to date t (i.e., the price of a zero-coupon bond starting at date t and ending at date T_j) we have:

$$Swap(t) = \delta \times N \times \sum_{j=1}^n (R^L(T_{j-1}, \delta) - F) B(t, T_j) \tag{I.B.7.3}$$

We now define the *swap rate* $S(t)$ to be the value of F such that $Swap(t) = 0$. That is,

$$\sum_{j=1}^n (R^L(T_{j-1}, \delta) - S(t)) B(t, T_j) = 0.$$

Thus (I.B.7.3) may be written

³ In our simplified notation, where $T_j - T_{j-1} = \delta$ and actual day counts are not used, of course all F_j are equal.

$$Swap(t) = \delta \times N \times \sum_{j=1}^n (S(t) - F) B(t, T_j) . \quad (I.B.7.4)$$

A payer swaption will only be exercised if $S(T_0) - F > 0$, indeed the pay-off is $[S(T_0) - F]^+$. It is like a call option with strike F . The underlying of the option is the *swap forward rate* computed at date t , denoted $F_S(t)$. Note that $F_S(T_0) = S(T_0)$ but for $t < T_0$ the swap forward rate is the current view of the swap rate at time t .

Thus, if we assume the usual lognormal dynamics for the swap forward rate we can apply the Black (1976) formula to value the swaption. The payer swaption pricing formula is then given by

$$Swaption(t) = \delta \times N \times \sum_{j=1}^n (F_S(t) \Phi(d) - F \Phi(d - \sigma_S \sqrt{T_0 - t})) B(t, T_j) \quad (I.B.7.5)$$

where σ_S is the volatility of $F_S(t)$, and

$$d = \frac{\ln\left(\frac{F_S(t)}{F}\right) + 0.5\sigma_S^2(T_0 - t)}{\sigma_S \sqrt{T_0 - t}} .$$

Note that the following pricing formula applies in the case of a receiver swaption:

$$Swaption(t) = \delta \times N \times \sum_{j=1}^n (-F_S(t) \Phi(-d) + F \Phi(-d + \sigma_S \sqrt{T_0 - t})) B(t, T_j) \quad (I.B.7.6)$$

I.B.7.5.2 Market Quotes

European swaption prices are expressed in terms of the implied volatility of the Black (1976) model. As for the cap/floor market, the market reference is the implied volatility curve as a function of the maturity of the option. As an illustration, we display here the volatility matrix for swaptions written on 1- to 10-year three-month Euribor swaps, as measured on a particular date t (see Figure I.B.7.2). The maturity of the swaptions ranges from 1 month to 10 years. The implied volatility quoted here is the mean of the bid and ask quotations. For example, the volatility of the one-year swaption written on the 10-year swap is 13.35%.

Figure I.B.7.2: Euribor swaption volatilities

<HELP> for explanation. DL19 Govt SWYV
 Enter 1 <Go> to save setting

Euro Volatility Swaption Implied
 (Bid/Ask/Mid) Page 2/3

Option Expiry	1 yr	2 yr	3 yr	4 yr	5 yr	7 yr	10 yr
1 mo	29.60 C	27.60 C	24.70 C	22.50 C	20.90 C	18.10 C	15.00 C
3 mo	25.85 C	24.70 C	22.60 C	20.90 C	18.65 C	16.95 C	14.10 C
6 mo	24.75 C	23.40 C	21.10 C	19.25 C	17.85 C	16.00 C	14.00 C
9 mo							
1 yr	22.65 C	20.60 C	18.80 C	17.40 C	16.30 C	14.25 C	13.35 C
2 yr	18.15 C	16.55 C	15.45 C	15.35 C	14.75 C	13.90 C	12.85 C
3 yr	16.90 C	15.30 C	14.60 C	14.10 C	13.70 C	13.20 C	12.50 C
4 yr	15.30 C	14.10 C	13.50 C	13.10 C	12.90 C	12.50 C	12.00 C
5 yr	14.20 C	13.40 C	12.80 C	12.40 C	12.30 C	12.00 C	11.60 C
7 yr	12.90 C	12.40 C	12.00 C	11.70 C	11.50 C	11.30 C	11.00 C
10 yr	12.00 C	11.60 C	11.30 C	11.10 C	10.90 C	10.80 C	10.60 C

Source: CMPN 20:59

<Menu> to select another ccy
 1 <Go> to save Bid/Ask/Mid
 2 <Go> to modify sources

Australia 61 2 9277 8600 Brazil 5511 3048 4500 Europe 44 20 7330 7500 Germany 49 69 920410
 Hong Kong 852 2977 6000 Japan 81 3 3201 8900 Singapore 65 212 1000 U.S. 1 212 318 2000 Copyright 2002 Bloomberg L.P.
 6356-711-0 12-Aug-02 23:14:39

Source: Bloomberg Eurozone

I.B.7.6 Uses of Swaptions

Like caps, floors and collars, swaptions are fixed-income securities designed to hedge the interest-rate risk. More precisely, a payer swaption can be used in two ways:

- It enables a firm to fix a maximum limit to its floating-rate debt.
- It enables an investor to transform fixed-rate assets into floating-rate assets to benefit from a rise in interest rates.

Example I.B.7.8

Consider a firm with a three-month Libor debt for the next five years and \$10 million principal amount. The swap rate for a three-month Libor swap with five-year maturity is 6%. The treasurer of this firm fears a rise in interest rates. Instead of entering a three-month Libor swap where he pays the fixed leg and receives the floating leg over a period of 5 years, he prefers to wait for an effective rise in rates. He contracts a payer swaption maturing in 6 months with a 6% strike rate and whose underlying swap has a 4.5-year maturity.

In 6 months, if the swap rate of the three-month Libor swap with a 4.5-year maturity is higher than 6%, the treasurer will exercise the swaption. On the other hand, if the swap rate is lower than 6%, he will give up the option, and if he wishes to modify the nature of its debt, he will enter a swap under better conditions at that time.

Symmetrically, a receiver option on swap can be used in two ways:

- It enables a firm to transform its fixed-rate debt into a floating-rate debt in a context of a decrease in interest rates.
- It enables an investor to protect a floating-rate investment.

Example I.B.7.9

A portfolio manager who has invested in a five-year maturity bond that delivers annually the one-year Libor, and expects a decrease in rates in 3 months, will typically contract a receiver swaption maturing in 3 months whose underlying swap has a 4.75-year maturity.

I.B.7.7 Summary

Caps, floors and swaptions are some of the most popular derivative instruments traded in the over-the-counter markets. They offer excellent opportunities for managing and speculating on interest-rate risk as they can be tailored to match a specific exposure or view. All three instruments are, in fact, options where the underlying asset is an interest rate. Accordingly, they can be priced by adapting the Black–Scholes–Merton stock option model.

References

Black, F (1976) The pricing of commodity contracts. *Journal of Financial Economics*, 3, pp. 167–179.

Martellini, L, Priaulet, S and Priaulet, P (2003) *Fixed-Income Securities: Valuation, Risk Management and Portfolio Strategies*. Chichester: Wiley.

I.B.8 Convertible Bonds

Izzy Nelken¹

I.B.8.1 Introduction

I.B.8.1.1 Convertibles – a definition

A convertible security is – quite simply – any security capable of being converted. Preferred stocks are one example. These either are, or eventually will become, convertible into common stocks or, in some cases, the cash value of common stocks, or into warrants that provide the owner the right to purchase common stocks at a specifically determinable price or schedule of prices. Other examples of convertible securities are bonds, debentures and notes. The definition includes ‘synthetic convertibles’ which may be created by combining separate securities that in combination possess the two principle characteristics of a true convertible security: an income stream or accretion, and the right to acquire equity securities or to participate in the future price performance of equity securities.

Convertibles are *hybrid* investment instruments, the most common of which are convertible bonds and preferred stocks. They are called ‘hybrid’ because most can be converted at the user’s choice into other investments, most often shares of common stock. Since they are a mixed breed, convertibles share the relative safety of fixed-income investments (bonds), while also being exposed to the underlying stock’s potential gains. Their value is correlated with the stocks into which they can be converted, yet their hybrid nature makes them less volatile.

Like fixed-income investments, convertibles pay interest and principal payments as well as dividends. In the case of a bond -- essentially a loan -- the company has to pay back the money with interest. For a preferred stock, which blends the characteristics of a bond and common share, it pays dividends and gives the investor a senior claim (over common stock) on a company’s assets in the event of a liquidation or sale.

According to www.kiplinger.com, convertible bonds combine the features of stocks and bonds in one investment. They are redeemable for a set number of shares of stock of the same company, or at a specified ratio – 25 shares of stock for each \$1000 in bonds, for example. If the price of the company’s stock rises, so will the price of its convertible bond, although not dollar for dollar. If the stock falls, the convertible will fall, too; but because it is a bond, its movement will also be affected by interest rates. Thus, fundamental or economic factors that are bad for stocks (e.g.,

¹ Super Computer Consulting, Inc., 3943 Bordeaux Drive, Northbrook, IL 60062, USA. www.supercc.com, www.optionsprofessor.com, izzy@supercc.com

recession) may be balanced by factors that are good for bonds (falling interest rates). If the stock rises high enough, you can exchange your bonds for stock and, if you wish, sell the stock and pocket the profits.

It sounds like the best of both worlds, but before you rush out to buy, consider a few more facts about convertibles:

- They are usually debentures, meaning they are not backed by the assets of the company.
- They are usually subordinated debentures, meaning other, unsubordinated debts will be paid off ahead of them in case of bankruptcy.
- Default risks aside, most convertibles are callable by the issuer at short notice. If the issuer exercises this right to buy back the convertible, the investor usually has a limited time (say, a month) in which to convert. This is known as a forced conversion. The issuer is unlikely to allow its stock price to climb very high above the conversion price before calling the issue; as a result, profit potential is limited.
- You pay a price for these hybrid securities. They pay less interest than you could get from the same company's bonds, and you can expect to pay a higher price to buy the convertible than its value as either a stock or a bond. That higher price is called the conversion premium or, in the parlance of Wall Street, 'water'. It represents the distance the stock price has to climb to reach the break-even point for converting the bond.
- This does not mean that convertible bonds are not worth considering. In a climate in which interest rates are declining and prospects look good for stock prices, as happens near the end of a recession, for example, carefully selected convertibles offer price potential and relatively small downside risk. But the interplay of interest rates, call provisions, conversion premiums and stock prices is so complex that it is safe to assume that individual investors get talked into buying convertibles by their brokers more often than they think of doing it themselves. In short, convertibles require more time and attention than most investors are willing to give them.

A related instrument to convertible bonds is 'convertible preferred shares'. These operate in a very similar fashion except for three main differences:

1. The par amount for a convertible bond is typically a round number (e.g., \$1000 or 100,000 Japanese yen).
2. Rather than a coupon, these typically pay a dividend. The dividend is set as a fixed amount, very similar to a fixed-rate coupon.
3. Convertible bonds rank senior to shares (but typically junior to other bonds) in the event of liquidation. Preferred shares are a different story. In most cases, they rank equally with the common stock.

Another related instrument is the ‘exchangeable bond’. This is convertible to common stock of another company (not the issuer).

I.B.8.1.2 Convertible Bond Market Size

As convertibles are traded over the counter (OTC), there are no exact figures, only estimates. Of the 500 companies in the S&P 500 index, roughly 100 issue convertibles. The largest ones are AIG, Berkshire Hathaway, UPS, MMM, General Motors, and Ford. At the time of writing the global market is estimated at \$610 billion: North America accounts for \$325 billion, Europe \$177 billion, Japan \$77 billion and Asia (other than Japan) \$31 billion. The market values of the convertible bond market in North America since 1987 and the value of new issues in the USA since 1995 are shown in Table I.B.8.1.

Table I.B.8.1: Convertible bond market size

Year	Market Value in North America (\$bn)	New Issuance in US (\$bn)
2003	305	89.4
2002	235	54.3
2001	252	106.8
2000	180	61.4
1999	195	41.3
1998	161	29.9
1997	158	33.8
1996	133	31
1995	128	17.2
1994	114	
1993	120	
1992	103	
1991	79	
1990	53	
1989	60	
1988	54	
1987	55	

Source: Convertbond.com

In 2003 the USA saw new issuance of some \$89.4 billion in convertibles, as compared with \$79.7 billion of equities. In many parts of the world, equity markets declined between March 2000 and December 2002. Investors have responded by shifting assets from equity (or growth) funds into fixed-income (or simply income) funds. Peter McInnes, head of structured capital markets at J.P. Morgan Australia, says the shift of funds from growth funds to income funds has halved the size of growth funds. Issuers have adapted to the shift in investor preference by issuing more convertible securities. In the first quarter of 2004, however, the relative popularity of convertibles versus equities was reversed. \$13.1 billion of convertibles were issued in the USA, compared with \$30.4 billion of equities. The popularity of convertibles in 2003 corresponded to very

favourable conditions: renewed strength in equity markets, stable interest rates and falling credit spreads all contributed to increased demand. However, in early 2004 these conditions reversed.

I.B.8.1.3 A Brief History

The first convertible bond issue was associated with J.J. Hill, the railroad magnate, in 1881. Hill believed that the market was ascribing too much risk to his rail project and needed an innovative way to secure long-term financing. Unwilling to sell stock until his planned expansion had reaped financial rewards, yet shut out from the traditional debt market, Hill issued a convertible bond. Obviously, since that time, there has been considerable growth and innovation in the convertible market, which is now over \$500 billion world-wide. However, convertibles still fulfil the same financing need as they did in Hill's day. Convertibles provide a way for companies whose stocks are volatile to access the debt market.

I.B.8.2 Characteristics of Convertibles

I.B.8.2.1 Relationship with Stock Price

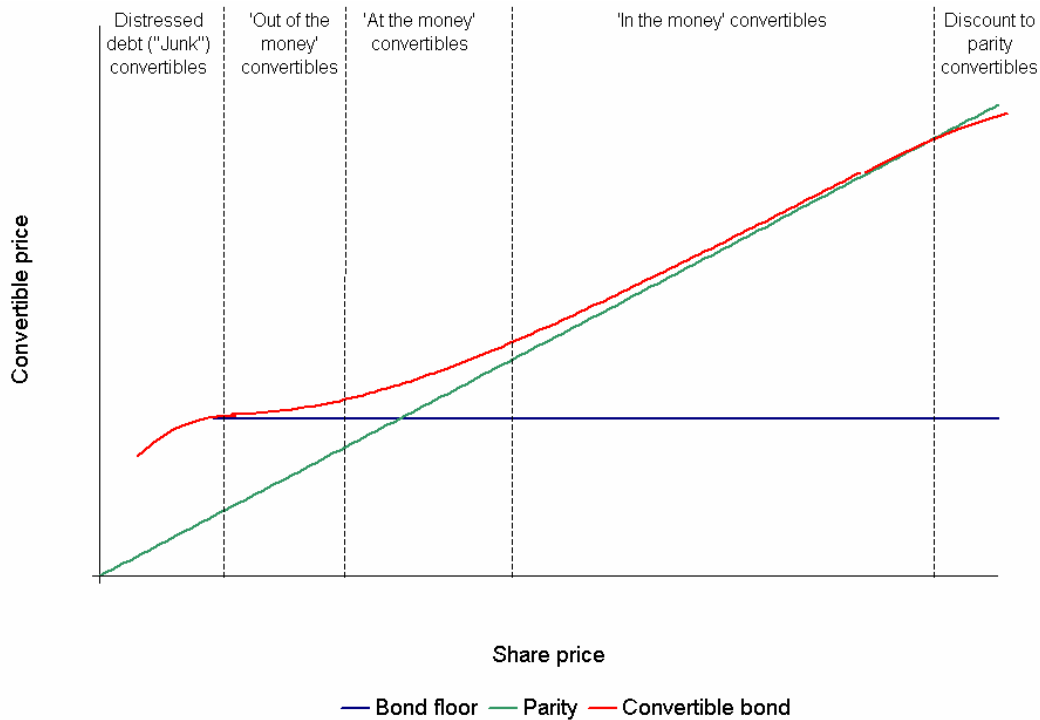
A convertible bond is a very interesting financial instrument. It will have elements of

- fixed income
- equity, and
- derivatives.

Obviously, we are speaking about a bond – which is a fixed-income instrument – but the bond has exposure to the equity markets, in the form of a derivative instrument. In Figure I.B.8.1 we qualitatively plot the value of a convertible bond against the underlying stock.

Obviously the convertible (red line) is worth more than a similar bond which is not convertible (blue line) because it has an additional right of conversion which will have a positive value. It is also worth more than the stock price times the conversion ratio (green line), otherwise investors could earn arbitrage profits by purchasing convertibles and immediately converting them to stock. We say, therefore, that the convertible bond's price must be above the 'bond floor' and also above the 'equity floor'.

Figure I.B.8.1: Relationship of the convertible to the underlying stock



It is customary to divide the stock price into five distinct regions. Convertible bonds have different characteristics and risk profiles in each region:

- *Distressed region*: When the stock price is very low, the company is in dire threat of bankruptcy. Credit spreads widen and the company paper is treated like ‘junk debt’. Further drops in the stock price will increase the credit spread and the risk of default. Owners of such a convertible bond are very sensitive to the credit risk of the issuer. They can hedge this risk through credit derivatives such as convertible bond asset swaps (see Chapter I.B.6).
- *Bond region* (‘out-of-the-money convertibles’): When the stock price takes values in this range, the company’s credit rating is deemed stable. The convertible bond behaves very much like a fixed-income instrument. Here, one is worried about interest-rate risk, which is typically measured using risk measures such as duration and convexity (see Chapter I.B.2).
- *Hybrid region* (‘at-the-money convertibles’): In this region, the bond behaves like a true hybrid instrument. The convertible bond’s price chart is convex.
- *Equity region* (‘in-the-money convertibles’): In this region, the convertible bond moves in line with the stock.

- *Very high stock prices* ('discount-to-parity convertibles'). The small discount of the convertible bond price to the stock price is normally because the bond is less liquid than the stock.

I.B.8.2.2 Call and Put Features

Convertible bonds and debentures are changeable, at the holder's option, into shares of the issuing corporation.² The conversion ratio is set out in the terms of the bond (or debenture). Typically it is spelled out in the prospectus. For example, a bond may be converted by the holder, at any time, to 50 shares of stock. If the stock is trading at \$17, the conversion feature provides a minimal floor on the price of the bond. It can always be converted to 50 shares and the shares could be sold for \$850.

In most new issues, the bonds or debentures are 'callable' by the issuer, usually at a premium to their issue price. In the case of convertible bonds, we distinguish between two types of call protection.

- Hard call protection: this means that the bond cannot be called until a certain date.
- Soft call protection: the bond could be called by the issuer but only if the share price exceeds a certain threshold. Typically, the threshold is well above the call price.

An investor can be certain that a convertible bond will not be called for the hard call protection period but cannot be certain about the soft call protection period. Whether the issuer is entitled to call the bond depends on the stock price.

For example, a convertible might have a hard call protection for two years. After that it is callable at par but only if the share price is above \$24. Assuming a conversion ratio of 50 shares per bond, the conversion value of the bond is at least $50 \times \$24 = \1200 when it is called. Of course, the issuer 'knows' that the bond holder will convert to shares (rather than accept par). Hence, when the issuer calls, it really forces investors to convert. Hence the name 'forced conversion'.

There is a concept of a 'notice period'. For example, the issuer must give 30 days' notice of intention to call the bond. In the meantime, investors may decide whether to accept the call or convert to shares.

Consider a different example in which the bond only had a hard call protection. After the hard call protection period expires, the issuer may call for par. If the issuer desires to force conversion, at what stock price should it call?

² Bonds that are convertible to shares of other corporations (not the issuer's) are called 'exchangeable bonds'.

Had it not been for the call notice, the issuer can call as soon as the stock price is larger than \$20. Assume that the stock price is \$20.01. A rational investor would convert to shares and receive \$1005.00 (rather than wait for the par amount – \$1000). However, the notice period complicates things a bit. It is quite possible that in the following 30 days the stock price will decline and the investors will choose to receive the cash. As the issuer typically wants to give out stock, it will call only when the stock is well above the \$20 threshold.

There are many other features of convertible bonds, including reset features and make-whole requirements, that are beyond the scope of this chapter.

I.B.8.2.3 Players in the Convertible Bond Market

We distinguish between several types of key players in this market:

- SEC/regulatory authority;
- issuers;
- investment banks;
- tax authority;
- rating agencies;
- investors.

Regulatory Authorities

The regulatory authorities require the maximum reasonable disclosure. For example, many jurisdictions require a prospectus to be filed that describes the bond in detail. For some securities a ‘short form prospectus’ is sufficient, while others require that a ‘long form’ prospectus be filed. Their objective is to protect the investing public without impeding capital flows. Other regulatory authorities set guidelines for banks in terms of regulatory risk capital. Such guidelines are concerned with the extent to which convertibles may be considered as capital, and prescribe the characteristics they must have to be treated as such. See Section I.B.8.3 on the implications of preferred shares on the capital structure.

Issuers

The issuer is concerned with obtaining the lowest possible funding costs. These include paying lower coupons, setting a high conversion premium³ and defining easy covenants⁴. Simplistically, we can say that from the point of view of the issuer:

³ Conversion premium is the difference between a bond price and the number of shares it can be converted to multiplied by the current share price. The issuer would like to sell the convertible at par but, at the same time, reduce its value. Allowing a conversion to a small number of shares is equivalent to setting a high conversion premium. For a detailed example, see Section I.B.8.5.

⁴ A covenant is a pledge or undertaking by an issuer to do certain things or avoid others. Usually, a covenant will be a ‘financial covenant’ which specifies that, for example, the issuer will maintain an interest coverage ratio over a certain

- Convertibles have a lower coupon than a non-convertible instrument.
- If it is not converted, the treasury saves in funding costs (as compared with a regular bond).
- If it is converted, the treasury sells shares at a premium

In some jurisdictions, the dividend on a preferred share is considered a coupon. The taxation authorities view them as an expense that is tax deductible. On the other hand, accounting treats these instruments as equity so they reduce the debt to equity ratios.

Convertibles are usually issued in the bond region (see Figure I.B.8.1). Typically, the conversion price is set to some 15–25% above the current stock price (this is done to encourage investors to hold on to the bond for a while). For example, assume that the current stock price is \$10 and convertible bond which can be converted to eight shares is sold for \$100. After several years, the investor converts, receiving eight shares. From the issuer's point of view, the investor has purchased eight shares for \$100, paying \$12.50 per share. Hence, the issuer seems to have sold shares at a premium.

Of course, the issue must be attractive enough to sell in the market, so recent convertible bond issues have put features and call notice periods that are designed to make the bond more attractive to the investors. One of the difficulties that an issuer faces is that of pricing a new deal. It must make the issue attractive enough to investors without 'giving away the farm'.

Investment Banks

The corporate services department of an investment bank is charged with assisting the issuer in bringing the issue to market. It advises the issuer on what terms will 'sell' in the marketplace. Like any other financial instrument, convertible bonds usually involve sales calls, road shows, etc. Some investment banks also act as market makers or brokers in this market. Others operate investment funds or hedge funds that specialize in convertible bonds or have investments or brokering relationships with such.

level or a leverage ratio (debt/equity) under a specific level. These ratios are meant to constrain the issuer to financial prudence. Covenants can also be 'non-financial' in nature, such as providing financial information to bondholders, protecting against the selling of assets, or changes of control, or making sure the assets of the company have adequate insurance (from www.finpipe.com).

Tax Authority

The taxation authority seeks to maximize revenue. There are numerous jurisdictional issues to consider here involving the taxation of both the issuer and the investor.⁵ For example, we have the following quote from David S. Lewis:⁶

“Convertible bonds and debentures are changeable, at the holder’s option, into shares of the issuer corporation. The rate of conversion is set out in the terms of the bonds or debentures. The holder of the debt instrument will normally exercise the right to convert shares when the value of those shares exceeds the value of the pure debt instrument. In some cases, the bonds or debentures are callable by the issuer, usually at a premium to their issue price. The exchange of the bond into shares takes place on a tax-free basis. See *Income Tax Act subsection 51(1)*. The holder’s adjusted cost base in the convertible bond is carried over to the shares, and the holder realizes a capital gain or loss when the shares are eventually sold.”

Rating Agencies

The rating agencies (S&P, Moody’s-KMV, Fitch, etc.) provide credit ratings to the issues. The bonds are assigned a letter code which corresponds to their credit quality (Table I.B.8.2).

Table I.B.8.2: Bond rating codes

Rating	S&P	Moody's
Highest quality	AAA	Aaa
High quality	AA	Aa
Upper medium quality	A	A
Medium grade	BBB	Baa
Somewhat speculative	BB	Ba
Low grade, speculative	B	B
Low grade, default possible	CCC	Caa
Low grade, partial recovery possible	CC	Ca
Default, recovery unlikely	C	C

⁵ For instance, from the corporation’s point of view, dividend income on preferred shares is tax deductible.

⁶ See <http://www.professionalreferrals.ca/article-495.html>

Investors

Investors have several methods of investing in convertible bonds. Some large institutional investors can invest in individual issues directly. Alternative investment vehicles are:

- (i) convertible bond funds;
- (ii) convertible arbitrage hedge funds.

We discuss each of these below.

I.B.8.2.4 Convertible Bond Funds

These funds typically invest in a portfolio of convertible bonds and preferred shares according to some investment philosophy.

The convertible bond fund has to build a portfolio of hybrid instruments according to a predefined mandate. Typically, the process follows three steps:

- Find an attractive equity.
- Find a cheap convertible related to that equity.
- Build a portfolio subject to risk constraints (e.g., do not deviate too much from some benchmark).

Table I.B.8.3: Strategy for convertible bond fund

		Equity Valuation		
		<i>Undervalued</i>	<i>Fair value</i>	<i>Overvalued</i>
Convertible Valuation	<i>Undervalued</i>	Buy signal		
	<i>Fair value</i>			
	<i>Overvalued</i>			Sell signal

Table I.B.8.3 sets out a typical trading strategy for a convertible bond fund. If the equity is undervalued and so is the convertible, a buy signal is issued. On the other hand, if the equity is overpriced and also the convertible is expensive, a sell signal is issued. In all other cases the bonds are purchased, sold or no action is taken according to the portfolio construction requirements.

Investment management of convertible bond funds typically fall into two types:

- (a) equity funds;
- (b) fixed-income funds

Many funds are a combination of these two distinct styles.

Equity funds find an attractive stock and then buy the convertible. They will do this in the hope that the stock will go up but with the protection offered by the convertible. They may be willing to pay slightly more than fair value for the convertible. The convertible may be so cheap that it is attractive compared to the common stock. So swap out of the stock and into the convert.

Fixed-income funds care much more about the bond value of the convertible, the coupon paid, etc. They desire some exposure to the equity markets. As an example, consider the American Express Convertible Bond Fund. They report the following (as of 31 March 2004).

Fund Objective: *The fund seeks a high total return through a combination of current income and capital appreciation. The fund seeks to achieve this objective by investing in a range of global convertible bonds.*

Investment Strategy: *The fund seeks to identify undervalued bonds with yields higher than the market as a whole. Sensitivity analysis is used to find convertible issues that are likely to capture more of the underlying stock's upside potential than downside price movement. The fund seeks to outperform its benchmark with less volatility.*

Risk and Return Profile: *For investors who seek moderate volatility similar to corporate bonds and return potential comparable to equities.*

Current Status [as at 31 March 2004]: *In March 2004, technology and healthcare – two sectors with typically higher betas – negatively impacted the portfolio. Late in the month, we took profits from five positions in technology. This took us from an overweight to underweight position in the tech sector and aligned the Fund's delta, currently at 50, with that of the Index. Consumer discretionary and energy, our overweight bets, continued to benefit performance. Credit quality regained its plus status in March and stands at BBB+.*

Net Asset Value Euro €16.42 or US \$20.12

Modified Duration 3.73 Years

Standard Deviation 4.12%

Beta 0.51

R² 0.85

Fund Size \$113.5 million

Quartile Ranking Last 12 Months: 2

Quartile Ranking Since Inception: 1

Sharpe Ratio: 0.50

Benchmark: Goldman Sachs Convertible Bond Fund

I.B.8.2.5 Convertible Arbitrage Hedge Funds

Hedge funds try to make money by purchasing or selling mispriced convertibles. They do *not* want to take a position in the equity. The typical core strategy involves establishing a long position in a convertible bond and a short position in the shares. The size of the short position is determined by the delta hedge ratio of the particular bond. They earn the coupon on the bond

but must pay out the dividend and the costs of borrowing short these shares. Thus a simple approximation for the positive carry on the trade can be given by:

$$\text{Carry} = \text{coupon} - \text{delta} \times (\text{dividend} + \text{short borrow costs}).$$

Typically, in addition to making money from a *mispricing* of the convertible, the hedge fund earns a positive carry, so many funds leverage up the trade many times over. Being actively delta-hedged, these funds are not exposed to small stock price movements.

Example I.B.8.1 (from <http://www.magnum.com/hedgefunds/convertibles.asp>)

Take a 5% convertible bond maturing in one year at \$1000, exchangeable into 100 shares of non-dividend-paying common stock currently trading at \$10 per share. An arbitrage strategy might hedge against this long convertible bond with a short position of 50 shares of underlying common stock at \$10 per share.

Return When No Change in Stock Price:

Interest payments on \$1000 convertible bond (5%)	\$50
Interest earned on \$500 short sale proceeds (5%)	\$25
Fees paid to lender of common stock (0.25% per annum)	(\$1.25)
<i>Net cash flow</i>	<i>\$73.75</i>
<i>Annual Return</i>	<i>7.4%</i>

Return When 25% Rise in Stock Price:

Gain on convertible bond	\$250
Loss on shorted stock (50 shares @ \$2.50/share)	(\$125)
Interest from convertible bond	\$50
Interest earned on short sale proceeds	\$25
Fees paid to lender of common stock	(\$1.25)
<i>Net trading gains and cash flow</i>	<i>\$198.75</i>
<i>Annual return</i>	<i>19.9%</i>

Return When 25% Fall in Stock Price:

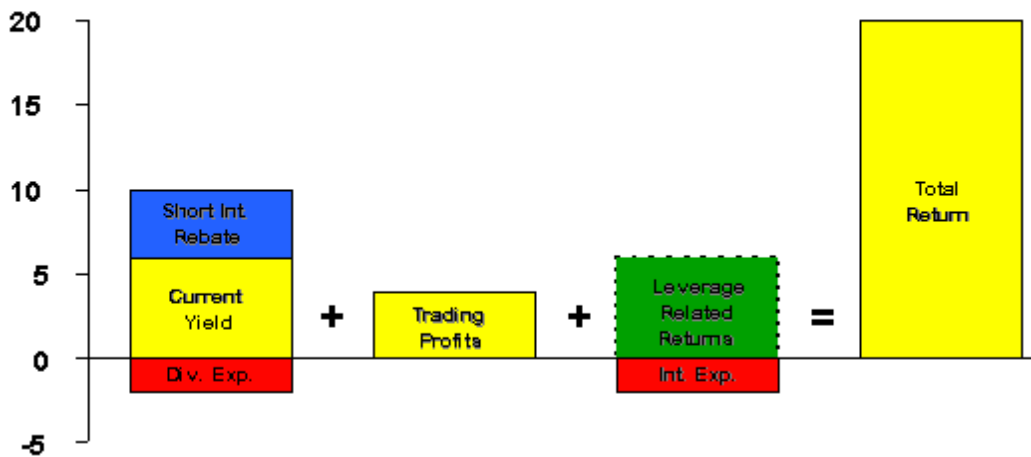
Loss on convertible bond ⁷	(\$80)
Gain on shorted stock (50 shares @ \$2.50/share)	\$125
Interest from convertible bond	\$ 50
Interest earned on short sale proceeds	\$ 25

⁷ Only falling as low as 'investment value'.

Fees paid to lender of common stock	(\$1.25)
Net cash flow	\$118.75
Annual Return	11.9%

The Magnum.com website shows the components of return from a convertible hedge fund strategy (Figure I.B.8.2). On the positive side are the current yield on the convertible, the short interest rebate and the trading profits. On the negative side there are the dividend expense and short borrow interest expense.

Figure I.B.8.2: Components of return on convertible arbitrage hedge fund



Some convertible hedge funds bill themselves as ‘convertible bond arbitrage’. Note that these strategies are not ‘arbitrage’ in the strict sense of the word. They involve risks from *several* sources, including the following:

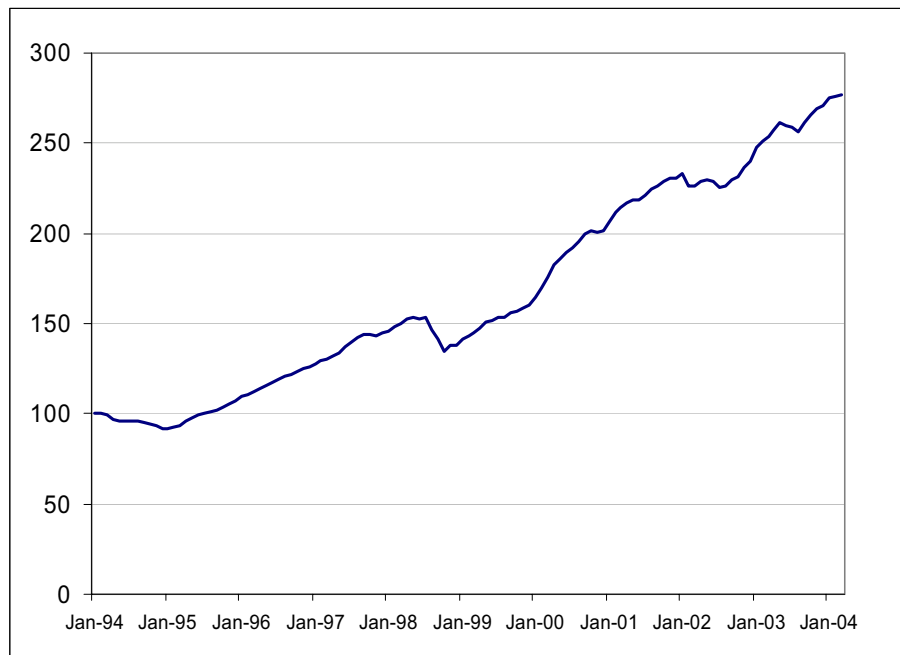
- *Interest-rate hikes*: these negatively affect the value of the bonds, and the shares might not react in a corresponding fashion.
- *Credit spread widening*: this will also hurt the bond. It is unclear whether or not the shares will move in a one-to-one relationship.
- *Share prices jumps*: these cannot be continuously hedged.
- *Implied volatility changes*: if the implied volatility in the bonds decreases, this will cause a drop in market value of the convertible.

A variety of techniques and financial products have been developed to assist hedge funds in protecting against these risks (at a price). These include: credit derivatives, variance and volatility derivatives (including swaps, futures and options), convertible bond asset swaps and so on.

The Credit Suisse First Boston / Tremont Convertible Hedge Fund index reports the historical returns shown in Figure I.B.8.3.

Although this chart of the index looks very impressive, as with any hedge fund index, this index suffers from many biases, including survivorship bias, liquidity constraints and reporting bias. Another difficulty is that one cannot invest in an index, only in individual hedge funds. So these statistics may be somewhat misleading. There is no guarantee that any single hedge fund manager could duplicate these impressive returns.

Figure I.B.8.3: Convertible arbitrage hedge fund index net asset value



I.B.8.3 Capital Structure Implications (for Banks)

Issuing preferred shares has profound implications on the capital structure of a bank. In the USA, the Federal Reserve allows perpetual preferred stock to be included in Tier 1 capital. Redeemable preferred stock and cumulative perpetual preference shares are typically included in Tier 2 capital.⁸ Thus preferred shares are treated as equity capital even though, as mentioned before, dividend income is a tax-deductible expense. This is ‘best of both worlds’ for the issuer.

Banks are very sensitive to their capital structure. Regulators impose certain liquidity ratios that must be maintained at all times (see Chapter III.0). The Basel Committee on Banking Supervision’s 1988 Capital Accord set capital requirements for major international banks from

⁸ See Section I.A.5.3 and http://www.federalreserve.gov/boarddocs/supmanual/bhcpr/s3_16_access.pdf for more information

G10 countries. The Accord requires these banks to ‘hold capital equal to at least 8% of a basket of assets measured in different ways according to their riskiness’. The capital referred to by the Accord is the sum of Tier 1 and Tier 2 capital. In addition, at least half the banks capital must be in Tier 1. Qualifying hybrid instruments can be included in Tier 2 capital, along with long-term subordinated debt and reserves. Further information about these requirements may be found at <http://www.bis.org>.

I.B.8.4 Mandatory Convertibles

Mandatory convertibles are instruments where the holder of the note *must* convert at the predefined ratio if he holds it to maturity. There are many types of mandatory convertible instruments, including *debt exchangeable to common stock* (DECS), also known as

- preferred equity participation securities,
- preferred redeemable increased dividend equity securities,
- mandatory adjustable redeemable convertible securities,
- stock appreciation income linked securities,
- threshold appreciation price securities,
- trust issued mandatory exchange securities, or
- trust automatic common exchange securities;

and *preferred equity redemption cumulative stock* (PERCS), also known as

- mandatory conversion premium dividend preferred stock,
- targeted growth enhanced term securities,
- yield enhanced stock,
- equity linked debt securities,
- performance equity-linked-redemption quarterly-pay securities,
- yield enhanced equity linked debt securities,
- and, in Europe, reverse convertibles.

DECS and PERCS may be callable by the issuer prior to maturity. If the issuer calls, the holder may convert.⁹

⁹ There are three types conversion rights of which may be granted: (i) no early conversion; (ii) early conversion allowed, but only at the lowest conversion ratio; and (iii) holder allowed to convert early according to the conversion schedule of the DECS.

Example I.B.8.2: Preferred equity participation securities

The payoff for preferred equity participation (PEP) securities is tied to the level of the share *except* when the share price falls in a certain range $[L, H]$: the payoff is

$$\begin{aligned} aS & \quad \text{if } S < L \\ C & \quad \text{if } L \leq S \leq H, \quad \text{where } C = aL = bH \\ bS & \quad \text{if } S > H. \end{aligned}$$

For instance, if on 31 December 2009 the share closes below \$50, the PEP is convertible to 0.8 shares. If it closes above \$50 and below \$80, the holder will receive \$40 in cash (or in shares) and if the share closes above \$80, the holder of the PEP will receive 0.5 shares.

Example I.B.8.3: Collar

The payoff for the collar is tied to the level of the share *only* when the share price falls in a certain range $[L, H]$: the pay-off is

$$\begin{aligned} L & \quad \text{if } S < L \\ aS & \quad \text{if } L \leq S \leq H \\ H & \quad \text{if } S > H. \end{aligned}$$

For example, this collar guarantees the receipt of at least \$80 in cash or shares. If on 31 December 2009 the share closes between \$80 and \$100, the holder will receive one share. If the share closes above \$100, the holder of the collar will only receive \$100. Owners of large blocks of shares who wish to protect the value of their shares, sometimes engage in “zero-cost collars”. This is done by purchasing a put option and selling a call option so that the premiums of both options are identical.

Example I.B.8.4: Other Mandatory Convertibles (e.g. double DECS)

One can create arbitrarily complex mandatory convertible structures. For instance, one can structure the payoff is as follows:

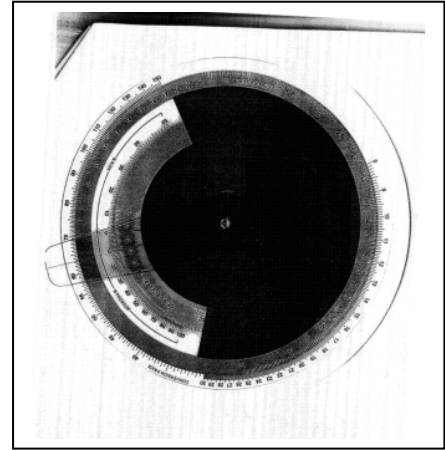
- If $S < 100$, receive one share.
- If $\$100 \leq S < \120 , receive \$100.
- If $\$120 \leq S < \144 , receive 0.8333 shares.
- If $\$144 \leq S < \150 , receive \$120.
- If $\$150 \leq S$, receive 0.8 shares.

Note that for mandatory convertibles, the coupon rate is set to the annual percentage rate received by the holder of the instrument. For example, DECS with a face value of \$54.125 paying an annual dividend of 7.875% or \$4.262 would be entered as an instrument with a face value of \$54.125 and a coupon of 7.875%.

I.B.8.5 Valuation and Risk Assessment

This ‘convertible bond slide rule’ was made in 1973. However, we do not explain how it worked here. We review only the simple methods for pricing and assessing the risk of convertible bonds. Many of these methods attempt to ‘score’ or ‘rank’ the convertible using relatively simple calculation. We can rank the convertible by:

- how high its price is over and above the common stock;
- how high its price is relative to a non-convertible bond.



(i) Minimum Value Approach

In this approach, the value V of the convertible is given by

$$V = \min(CV, BV)$$

where CV is the conversion value if converted immediately, and BV is the straight bond value.

This approach may be simple but it is not accurate. Rather than an approximation, it only gives a lower bound to the value of the convertible. In addition, when calculating the straight bond value, this approach encounters problems. When an issuer calls a straight bond, the investor must accept the call price. If a convertible bond is called, the investor can convert into shares.

(ii) The Market Conversion Premium

In this method, we attempt to find the mark-up involved in purchasing a convertible rather than buying the common stock. From the investor’s point of view, the lower the conversion premium, the better, all other things being equal. This is calculated with some simple formulae:

$$\text{Market conversion price} = \text{Market price of the convertible} / \text{conversion number}$$

$$\text{Market conversion premium per share} = \text{Market conversion price} - \text{current stock price}$$

$$\text{Market conversion premium ratio} = \text{Market conversion premium per share} / \text{current stock price}$$

Example I.B.8.5: Calculating the conversion premium

Suppose:

Convertible bond price	\$950
Face value	\$1000
Convertible to:	50 shares
Market price of share	\$17

Then:

$$\begin{aligned} \text{Market conversion price} &= \$950 / 50 = \$19 \\ \text{Market conversion premium per share} &= \$19 - 17\$ = \$2 \\ \text{Market conversion premium ratio} &= \$2 / \$17 = 11.8\% \end{aligned}$$

In essence, the investor pays a premium of 11.8% when purchasing the bond.

(iii) Breakeven Period Analysis

This analysis compares two alternatives: owning shares and owning convertibles. It first computes the ongoing *yield advantage* (i.e., coupon rate minus dividend rate), and then calculates the time required by the yield differential to fully compensate the initial conversion premium. The (favourable) *income differential* per share is given by:

$$(\text{Coupon interest from bond} - (\text{conversion ratio} \times \text{annual dividends})) / \text{conversion ratio}$$

Then the *breakeven time* is

$$\text{Market conversion premium per share} / \text{favourable income differential per share.}$$

It is called the ‘breakeven time’ because it gives the time (in years) for the enhanced income from the bond to compensate for the conversion premium. All others things being equal, the investor would prefer to own bonds with short breakeven times. For more details, see Fabozzi (2000).

Example I.B.8.6: Calculating the breakeven period

For the bond in Example I.B.8.5, suppose also:

Interest from bond	10%
Dividend per share	is \$1 per year

Then

$$\begin{aligned} \text{Coupon interest from bond} &= \$100 \\ (\text{conversion ratio} \times \text{annual dividends}) &= \$50 \\ \text{Favourable income differential per share} &= (\$100 - \$50) / 50 = \$1 \\ \text{Breakeven time} &= \$2 / \$1 = 2 \end{aligned}$$

Breakeven period analysis only serves as a sorting mechanism for convertible securities. Convertible securities with a breakeven period of less than 3 years are generally deemed as acceptable investments. All other things being equal, investors would desire a short breakeven period. This method does not attempt to value the options embedded in the convertible security and may therefore give a misleading signal as to its true value. Accordingly, none of the uncertain risk factors that relate to option valuation – future dividend payments, option time value and stock price volatility – are considered in the model.

(iv) Discount Cash-flow Analysis

Similar to breakeven period analysis, this analysis is also based on yield advantage but goes further and projects future cash flow. That is, it projects the coupon income minus dividend income, with the possibility of incorporating dividend growth. The net present value of the projected cash flows is added to the current stock price. The sum is then multiplied by the conversion ratio to arrive at a ‘theoretical’ convertible value.

Example I.B.8.7: Discounted cash flow analysis

In Examples I.B.8.5 and I.B.8.6, the stock price was \$17. The current income advantage of the convertible over the stock is \$1. For simplicity’s sake, assume a dividend growth of 10%, a flat interest rate of 5% and a two-year horizon. Then, under annual compounding, we have the following table:

Time	Coupon	Dividend	Coupon- Dividend	PVCF	Per Share
0	\$ 100.00	\$ 50.00	\$ 50.00	\$ 50.00	\$ 1.00
1	\$ 100.00	\$ 55.00	\$ 45.00	\$ 42.86	\$ 0.86
2	\$ 100.00	\$ 60.50	\$ 39.50	\$ 35.83	\$ 0.72
				Total	\$ 2.57

The ‘fair value’ of the bond would be $(\$17 + \$2.57) \times 50 = \$978.50$.

Again, this method ignores the option time value and underlying stock price volatility. It is a better method than breakeven period analysis because it actually values a convertible security. It is more applicable when the convertible is being evaluated as a debt rather than equity instrument.

(v) Investment Premium

A typical measure of the investment premium of a convertible is the *premium over straight value*, which is defined as:

$$(\text{Market price of a convertible} / \text{Straight value}) - 1$$

This simply gives a measure of the price we are paying for the convertible, over and above a similar non-convertible bond. All other things being equal, investors would desire a low investment premium.

Note that there is a salient difficulty here. Assume that the convertible bond has a soft call feature. Should we consider the same feature in a similar straight bond? The call option of a straight bond typically has nothing to do with the value of the share price.

Example I.B.8.8: Premium over straight value

For the bond in our example above, suppose there is a straight bond of the same issuer currently trading at \$788. Then the convertibles premium over this value is $\$950 / \$788 - 1 = 21\%$.

(vi) Synthetic Approach (Bond + Equity Option)

This more advanced approach to valuation of convertible securities assumes the value of a convertible is equal to the sum of two components: a straight bond of the same coupon and maturity, and a call option on the stock with the same exercise price. The bond component, in the absence of a comparable bond with the same coupon and maturity, can be valued by discounting all future coupons and the maturing principal by the appropriate risk-adjusted discount curve. The option will be valued independently with a long-term option model; see Fabozzi (2000).

Example I.B.8.9: Simple synthetic valuation

In this method, we will compute the price of the bond as:

$$\$788 + 50 \times (\text{Call option with a strike of } \$20)$$

The price of the call option depends on several factors. Firstly, the volatility of the share in question must be considered. Another problem is that of determining the expiry date of the option. Is it the final maturity of the bond? Is it the first call date? The second date?

Consider a specific option. Differing expiration assumptions will result in different option prices and therefore, different bond prices:

Maturity (years)	Option Premium	Convertible Price
1	\$ 1.21	\$ 848.50
2	\$ 2.25	\$ 900.55
5	\$ 4.48	\$ 1,012.04
10	\$ 6.80	\$ 1,127.90

Obviously, this naïve method is highly dependent on the choice of expiry date.

(vii) Modern Methods

A closer look at the conversion feature reveals that the equality of a convertible to the sum of a straight bond and a call option on stock is a misnomer. There is a major difference between a straight freely traded equity option which is exercisable by paying a known fixed exercise price, and a convertible's converting feature which is exercisable by turning in the value of a bond, whose value at any point in time depends on the value of the forgone coupons relative to the current yield curve. Valuation is further complicated by the fact that both the issuer and the investor typically have an option. The former has an option to buy the convertible and the latter has an option to convert into shares. The interaction between these rights creates complexity for valuation purposes.

Neither breakeven period analysis nor discount cash-flow analysis capture the full range of call or put features properly. By contrast, the synthetic approach handles the option feature separately, to account for the option time value and volatility. Most recent advances in convertible bond pricing models use an advanced synthetic valuation approach in a multi-factor framework, where multiple risk factors include stock price, its volatility, interest rates, default rates and, for cross-currency convertible securities, foreign exchange rates. The correlation between these factors can be a key issue to address, but is beyond the scope of the PRM syllabus. Students wishing to find out more about these methods are referred to Nelken and Cheung (1994) and Nelken (1999). In these papers, we discuss the usage of quadronary tress to value convertibles. Recently, there has been a trend towards valuation methods that rely on solutions to Partial differential equations. Tavella and Randall (2000) is an excellent source.

I.B.8.6 Summary

Convertible bonds and preferred shares are an important and growing asset class. These are complicated to structure, analyze and trade. In this chapter we have described the market for convertible bonds and the users of these instruments. We have described how the various entities involved may have differing interests. We have also covered some elementary valuation techniques. The more advanced valuation and risk management techniques are beyond the scope of this chapter. It is clear that advances will continue in this field. It is expected that the structures will get more complicated and also the valuation and risk management techniques. In the not so distant future, our best valuation methods will look as arcane as the circular slide rule from 1973.

References

Fabozzi, F J (2000) *The Handbook of Fixed Income Securities* (6th edition). Englewood Cliffs, NJ, Prentice Hall.

Nelken, I (1999) Japanese reset convertible bonds and other advanced issues in convertible bonds. *Financial Engineering News*, January. Available from either www.supercc.com or <http://www.fenews.com/fen8/japanese.html>

Nelken, I, and Cheung, W (1994) Costing the converts. *Risk Magazine*, 7(7), pp. 47–49.

Tavella, D, and Randall, C (2000) *Pricing Financial Instruments – The Finite Difference Method*. New York: Wiley.

I.B.9 Simple Exotics

Catriona March¹

I.B.9.1 Introduction

There has been spectacular growth in the derivative markets over the last three decades. Not only has there been an explosion in volumes but also a proliferation in the variety of products available. In particular, exotic options are now commonly traded throughout the equity, commodity, foreign exchange and interest-rate markets.

A standard European call (put) option allows the holder to buy (sell) the underlying asset on the expiry date for an agreed strike price. Because this is seen as the most basic option contract, it is often referred to as a *vanilla* option. An *exotic* option is an option whose payoff is non-standard in some way. Exotic options are traded as standalone products in both over-the-counter and exchange-traded markets. They are also attached to other instruments to enhance the appeal of the whole package, such as corporate bonds with exotic features. Some exotics are regularly traded; others have been introduced but continue to be unusual in practice. Textbooks giving catalogues of exotic options include Haug (1997) and Zhang (1997).

The purpose of this chapter is to give an introduction to the different types of exotic options traded in the financial markets. There are too many possible variations in exotic payoffs to cover all of them here. We describe the varieties that are most prevalent in the market, as well as simple exotics whose features are likely to be used in constructing other more complex products. Their payoffs are defined; motivations for their use and difficulties in their management are also discussed.

Since the 1970s, the Black–Scholes–Merton model has been widely used in the financial markets. Originally introduced for the pricing of vanilla options (see Chapter I.A.8) it is possible within the framework of this model to derive analytical formulae for the prices of many types of exotic options. Analytical pricing is important for efficient valuation and risk management. Although space does not permit us to reproduce these formulae here, many can be found in Haug (1997) or Zhang (1997), with references to their original derivations. Other exotics, especially those with more complex payoffs, can only be priced using numerical methods. Pricing exotics in more general models which allow for stochastic volatility or discontinuous asset prices also necessitates

¹ Applied Finance Centre, Macquarie University, Sydney, Australia.

numerical methods. A brief overview of techniques commonly used for valuing exotic options is given in Section I.B.9.14.

I.B.9.2 A Short History

While options have been traded throughout history, with olive contracts in ancient Greece and tulip options in the market ‘bubble’ in Holland during the 1630s often cited as examples, modern derivatives markets as we know them today have developed since the 1960s. This development and growth in derivatives markets has been driven by changing attitudes towards risk. Since the 1970s, the markets have seen floating exchange rates, periods of historically high oil prices and interest rates, as well as stock market crashes. The higher volatility of financial asset prices and the difficulties this brought to corporate and institutional management resulted in a new awareness of the necessity for financial risk management.

Over the same period there were vast improvements in technology, both the discovery of theoretical pricing techniques and the availability of more powerful computer hardware. This enabled banks and other market-makers to develop new products and manage portfolios containing them within a reasonable period of time, and so satisfy the rising demand for sophisticated risk management tools. Exotic options in particular could be customised to suit an individual’s hedging needs or speculative views.

During the 1960s, options were valued on an *ad hoc* basis. Over-the-counter barrier options were traded during this period. The Black–Scholes–Merton (BSM) model (Black and Scholes, 1973; Merton, 1973) provided a rigorous formula for pricing standard European options. More importantly, its concept of risk-neutral valuation showed that these options could be hedged perfectly within the framework of the model, giving option providers a method for managing their own risk. The theory of exotics valuation also dates from this landmark, with the formula for a down-and-out call option appearing in Merton (1973).

The 1970s and 1980s saw the gradual discovery of pricing techniques and formulae for particular exotics mainly within the BSM model. For example, using Monte Carlo simulation for pricing derivative contracts, in particular average rate and other path-dependent options, was pioneered in Boyle (1977); the analytical formula for exchange options appeared in Margrabe (1978); lookback options were first valued in Garman (1989).

The next major theoretical breakthrough for exotics option pricing was the ‘Fundamental Theorem of Option Pricing’ in Harrison and Pliska (1991). Within a market model which is arbitrage-free and which has a unique risk-neutral probability measure, any option payoff can be

replicated by a dynamic portfolio of the market assets and a risk-free borrowing; then the value of the option is given by its discounted expected value under the risk-neutral measure. This provided a general methodology for valuing any exotic payoff within the BSM model, the BSM model generalised to multiple assets, and more general models, opening the way to price any exotic option.

By the early 1990s, the markets for standard options had become extremely competitive and banks looked to exotic options as a way of increasing profits. Clients were also familiar with standard options and willing to try something new to optimise their hedging. The next few years saw an explosion in the development and marketing of new exotic products, as is evidenced in Jarrow (1995), a compilation of *Risk Magazine* articles from the period. Basket, contingent premium, rainbow, quanto, spread and many other types of exotic options became more commonplace. The term ‘exotic option’ was coined by Mark Rubinstein in a collection of articles circulated in the early 1990s.²

Towards the middle of the 1990s, several large corporate losses involving the use of derivatives for hedging or trading, including Metallgesellschaft in 1993, Procter and Gamble in 1994 and Orange County in 1995, led to a cooling of the market for exotic products. The need to hedge exposure to risky financial assets still remained, but there was a move towards cheaper, more manageable exotics, by both product providers and their clients. For example, average rate options and barrier options are cheaper than equivalent vanilla options. Average rate options are now so widespread in the commodity markets that they have almost come to be viewed as ‘standard’ rather than ‘exotic’ by market participants. In the currency market, exotic options are estimated³ to be approximately 10% of options traded; of these 9% are barrier options, while all other exotics make up only 1%; hence the term ‘exotic’ is sometimes used synonymously for ‘barrier’ option by foreign exchange options traders.

In the wake of the stock market crashes of 1987 and 1997, but especially since the late 1990s, the market has become aware of the limitations of the BSM model. The BSM formula continues to be used for vanilla options but a different volatility is used for each strike price and maturity, leading to the volatility smile typical of most option markets. Sophisticated models that may include stochastic interest rates,⁴ stochastic volatility and jumps in the asset price or the volatility are being used more often by market participants. Valuing exotic options within such models is more challenging than in the BSM model, but in recent years these models have begun to be applied to exotics, as in the compilation of papers in Lipton (2003). However, there is as yet no

² See Rubinstein (1992).

³ U. Wystuo, Bachelier Society presentation, 2002.

⁴ The term ‘stochastic’ means that the variable in question, say interest rate or volatility, changes in a random fashion.

consensus on a new paradigm to replace the BSM model; see, for example, Ayache *et al.* (2004). In particular, many models can fit the market prices of vanilla options but give quite different prices for exotic options and also quite different hedge ratios for vanilla options. This is currently the major outstanding problem to be resolved in the theory and market practice of exotic options trading.

I.B.9.3 Classifying Exotics

There are countless ways in which the payoff of an exotic option may be different from that of a standard European option. It is worthwhile making the following distinctions:

- The payoff may depend on a *single asset* price or on *multiple asset* prices.
- The payoff may depend on the asset price at different points in time before expiry (*path-dependent*) or only on the value at expiry (*path-independent*).
- Path dependence may be on *discrete* points of time or *continuous* sections of time.
- The underlying asset may be a derivative (an option on another option); these are sometimes called *second-order contracts*.
- The holder may be able to make a *decision* which will affect the option's payoff or its timing.

In the table below, the different exotic options discussed in later sections are organised according to these criteria.

	Single-asset	Multi-asset
Path-independent	Cash-or-nothing Asset-or-nothing Truncated Contingent premium Second-order: Compound, instalment Chooser Extendible	Two-colour rainbow Spread Outperformance: - Quotient - Exchange - Min or max of two - Relative performance Quanto Basket
Path-dependent	Barrier: - Single, double - Full-life, partial - Parisian - One-touch, no-touch Ladder Lookback, hindsight Average options: Average rate, average strike Double average rate Forward start, cliquet Strike reset Decisions: American, Bermudan, shout	Outside barrier Quanto barrier Basket lookback Basket average Quanto average American basket

In the bottom right of the table we can see that other exotics can be constructed by combining different features. Certainly complex composite products can be found in over-the-counter markets. The more complex the payoff, the more likely that a numerical technique will be necessary for its valuation. This may mean that pricing them quickly enough for risk management may be difficult. Also they may be too risky for a market-maker to provide because they are too difficult to hedge; they may be too complex for the buyer's needs, or simply not make business sense.

I.B.9.4 Notation

An exotic option is defined by the payoff it gives. In writing the payoffs below, we shall use the following notation:

T = the time until the option expires

K = the option strike price

$S(t)$ = the asset price at time t

$V(t)$ = the option price at time t

Write x^+ for the positive part of a number x , that is:

$$x^+ = \max(x, 0)$$

The payoff at expiry of a standard call option is:

$$V(T) = [S(T) - K]^+$$

The payoff at expiry of a standard put option is:

$$V(T) = [K - S(T)]^+$$

In fact both of these payoffs can be written as:

$$V(T) = [\lambda(S(T) - K)]^+$$

where we have introduced a parameter λ so that:

$\lambda = 1$ for the call option, and

$\lambda = -1$ for the put option.

An *indicator function* is a function taking the value 1 if some condition is satisfied and 0 otherwise:

$$\mathbf{I}_A = \begin{cases} 1 & \text{if condition } A \text{ is true,} \\ 0 & \text{otherwise.} \end{cases}$$

The *intrinsic value* of a standard option is its value if exercised at the current asset price $S(t)$, that is

$$[\lambda(S(t) - K)]^+.$$

If this amount is greater than zero, the option is said to be *in-the-money*. Clearly this is the case when:

$$S(t) > K \text{ for a call,}$$

$$S(t) < K \text{ for a put.}$$

The option is said to be *out-of-the-money* when:

$$S(t) < K \text{ for a call,}$$

$$S(t) > K \text{ for a put.}$$

Its intrinsic value is zero. If $S(t) = K$, the option is the *at-the-money*. If $S(t)$ equals the current forward asset price to the option expiry date, it is referred to as *at-the-money-forward*.

I.B.9.5 Digital Options

A *digital* option is one that pays either a fixed quantity if some condition is met or nothing. They are also called as *binary* options. Digital options often occur embedded in other structures.

I.B.9.5.1 Cash-or-Nothing Options

The simplest exotic payoff is a *cash-or-nothing* option, because it either pays 1 unit of currency (times a contract face value) or nothing.

The payoff for a ‘call’ is:

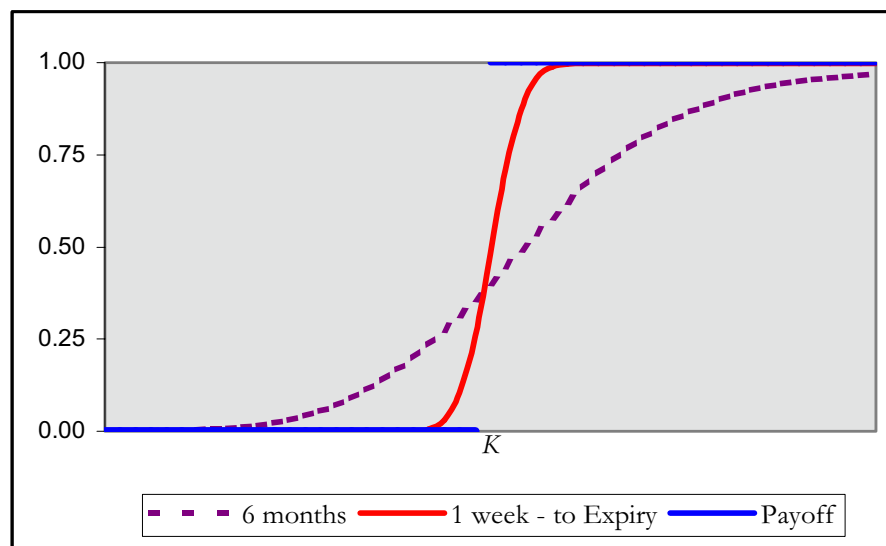
$$V(T) = \mathbf{I}_{\{S(T) \geq K\}} = \begin{cases} 1 & \text{if } S(T) \geq K \\ 0 & \text{if } S(T) < K. \end{cases}$$

The ‘put’ payoff is:

$$V(T) = \mathbf{I}_{\{S(T) \leq K\}} = \begin{cases} 0 & \text{if } S(T) > K \\ 1 & \text{if } S(T) \leq K. \end{cases}$$

They are called ‘call’ and ‘put’ only by analogy with standard options; at expiry a cash flow may occur but no asset is actually bought or sold. Figure I.B.9.1 shows the value of a call.

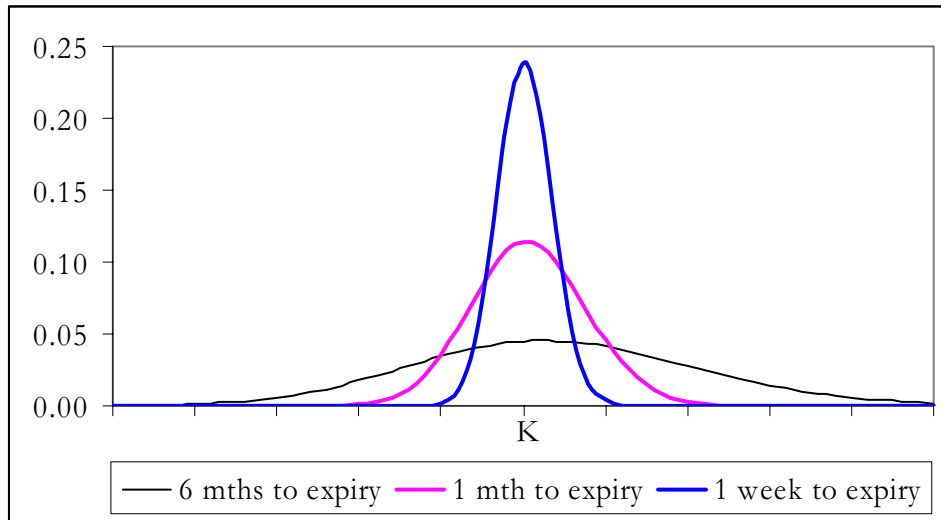
Figure I.B.9.1: Payoff for cash-or-nothing call



The price of a cash-or-nothing option is the (risk-neutral) probability that the option condition will be met, that is, that $S(T) \geq K$ for a call or $S(T) \leq K$ for a put, discounted to present value.

Despite their simplicity, the discontinuous nature of the payoff means that cash-or-nothing options are difficult to hedge, especially as expiry approaches. The option value changes rapidly from 0 to 1; correspondingly, the delta is large if the asset price is near the strike price and close to expiry. As the delta rapidly increases then decreases in this situation, the gamma will be large and will change sign as the asset price goes through the strike price.⁵

Figure I.B.9.2: Cash-or-Nothing Call Delta



I.B.9.5.2 Asset-or-Nothing Options

The holder of an *asset-or-nothing* call (put) option receives the asset if the asset price is greater (less) than the strike price at expiry. So the payoff for a call is

$$V(T) = S(T) \mathbf{I}_{\{S(T) \geq K\}} = \begin{cases} S(T) & \text{if } S(T) \geq K \\ 0 & \text{if } S(T) < K. \end{cases}$$

The payoff for a put is

$$V(T) = S(T) \mathbf{I}_{\{S(T) \leq K\}} = \begin{cases} 0 & \text{if } S(T) > K \\ S(T) & \text{if } S(T) \leq K. \end{cases}$$

Notice the vanilla option payoff can be rearranged as

$$[\lambda(S(T) - K)]^+ = \lambda S(T) \mathbf{I}_{\{\lambda(S(T)-K) \geq 0\}} - \lambda K \mathbf{I}_{\{\lambda(S(T)-K) \geq 0\}}.$$

This means that:

Payoff of vanilla option

$$= \lambda \times \text{Payoff of asset-or-nothing option} - \lambda \times \text{Strike} \times \text{Payoff of cash-or-nothing option}$$

⁵ See Chapter I.A.8 for the definition of delta and gamma.

At the expiry of a vanilla option, we exchange the strike price in cash for the asset, when the exercise condition is met. To avoid arbitrage, the same relationship must hold at any time before expiry. Alternatively, being long an asset-or-nothing option is the same as being long the equivalent⁶ vanilla option and short the equivalent cash-or-nothing option.

I.B.9.5.3 Vanillas and Digitals as Building Blocks

Common option strategies consisting of combinations of standard options include the following:

- a *straddle* is a call plus a put with the same expiry and the same strike price;
- a *strangle* is a call strike K_1 plus a put strike K_2 , where $K_1 > K_2$, with the same expiry;
- a *risk-reversal* is a call strike K_1 minus a put strike K_2 , where $K_1 > K_2$, with the same expiry;
- a *butterfly* is a risk-reversal minus a straddle.

See Chapter I.B.5 for more details on these strategies. When trading a strangle or risk-reversal, the call and put are generally out-of-the-money; the straddle component of a butterfly is generally at-the-money. In fact, any piecewise linear payoff can be constructed from a portfolio of standard and digital options at the appropriate strike prices. Some seemingly ‘exotic’ option payoffs turn out to be a linear combination of standard and digital options. The price of the package is just the sum of the prices of its components; likewise for its delta and other ‘Greeks’.

For instance, in a *truncated call option*, to reduce the cost of a standard call option, the holder agrees to forgo the payoff beyond a certain point. The payoff of such an option, with strike K and cut-off level $H > K$, would be

$$V(T) = \begin{cases} 0 & \text{if } S(T) \geq H \\ S(T) - K & \text{if } K < S(T) < H \\ 0 & \text{if } S(T) \leq K. \end{cases}$$

Writing $V^c(t, K)$ for the price at time t of a standard call with strike K and $V^d(t, K)$ for the price at time t of a digital call with strike K , the truncated option payoff can be constructed from

$$V(T) = V^c(T, K) - V^c(T, H) - (H - K)V^d(T, H).$$

The price now at time $t < T$ is given by the same relationship:

$$V(t) = V^c(t, K) - V^c(t, H) - (H - K)V^d(t, H).$$

⁶ That is, with the same strike price and expiry.

A *bet option* pays a fixed amount if the asset price expires between two levels, say L and H where $L < H$, so in fact is just the difference between two cash-or-nothing options:

$$V(t) = V^D(t, L) - V^D(t, H).$$

I.B.9.5.4 Contingent Premium Options

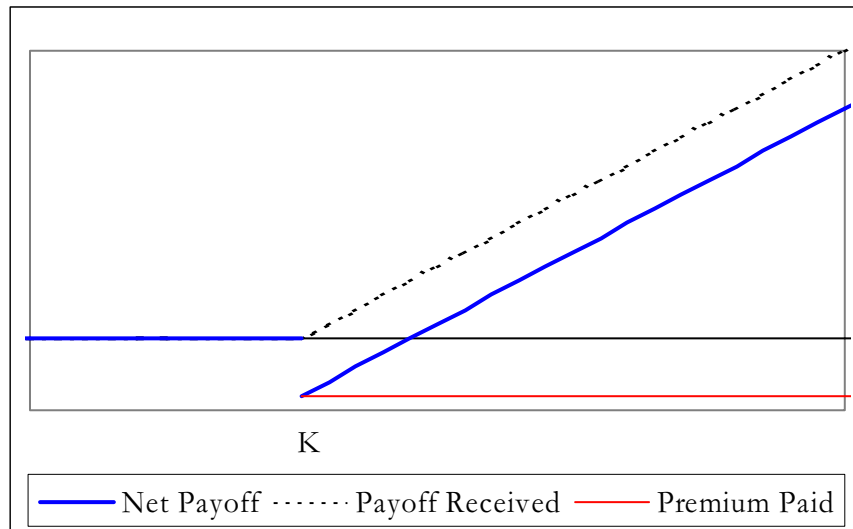
These are also called *pay-later options* because, rather than paying an option premium at the inception of the contract, the premium is only paid at expiry *if the option is in-the-money*; no premium is paid if the option expires out-of-the-money. Suppose the agreed premium is an amount α ; then the payoff for a call is

$$V(T) = \begin{cases} S(T) - K - \alpha & \text{if } S(T) \geq K \\ 0 & \text{if } S(T) < K; \end{cases}$$

and for a put

$$V(T) = \begin{cases} 0 & \text{if } S(T) > K \\ K - S(T) - \alpha & \text{if } S(T) \leq K. \end{cases}$$

Figure I.B.9.3: Contingent Premium Call Payoff



Clearly once the amount α is set the pay-offs are related as follows:

$$\text{Pay-later option} = \text{Vanilla option} - \alpha \times \text{Cash-or-nothing option}.$$

Often α is chosen so that the structure is initially zero-cost. In this case we have:

$$\alpha = \text{Vanilla option price} / \text{Cash-or-nothing option price}.$$

Because the price of a cash-or-nothing option is the discounted probability that it is in-the-money at expiry, this shows that the premium to be paid later will *always be bigger* than the premium of the equivalent vanilla option. For instance, for a strike price that is at-the-money-forward, the cash-or-nothing price is approximately 0.5; the premium to be paid later for the contingent premium option will be approximately twice the price you would pay now for an at-the-money-forward vanilla option.

In this way the holder of the option *does* pay for the extra advantage of not paying a premium if the option expires worthless. If the purpose of purchasing an option is to buy insurance against a worst-case scenario happening, the holder may not want to pay an increased premium when the worst-case scenario actually eventuates.

I.B.9.5.5 Range Notes

In a *range note* interest only accrues on days when the index rate, for example LIBOR, falls within some range. So it is really a note with an attached series of digital options, one for each day of the life of the note. The rate at which interest accrues is set at a spread over the corresponding ordinary note rate, to compensate for the probability that no interest will be paid on some days. Such notes are sometimes called *accrual* or *accumulation* notes, also *fairway* or *corridor* bonds.

In an *index range note*, whether interest accrues or not depends on an equity index or exchange rate. Again the idea is to gain extra yield when an investor has confidence in the view that the index will remain within a certain range.

In another variation, the index may have to stay within the range for the whole period to receive a rate enhancement. In this case the attached option is a digital barrier option (Section I.B.9.12) rather than a series of expiry-only digitals.

I.B.9.5.6 Managing Digital Options

Although cash-or-nothing options are probably the easiest type of option to price in any model, because of their discontinuous payoff they are also one of the hardest to manage in practice.

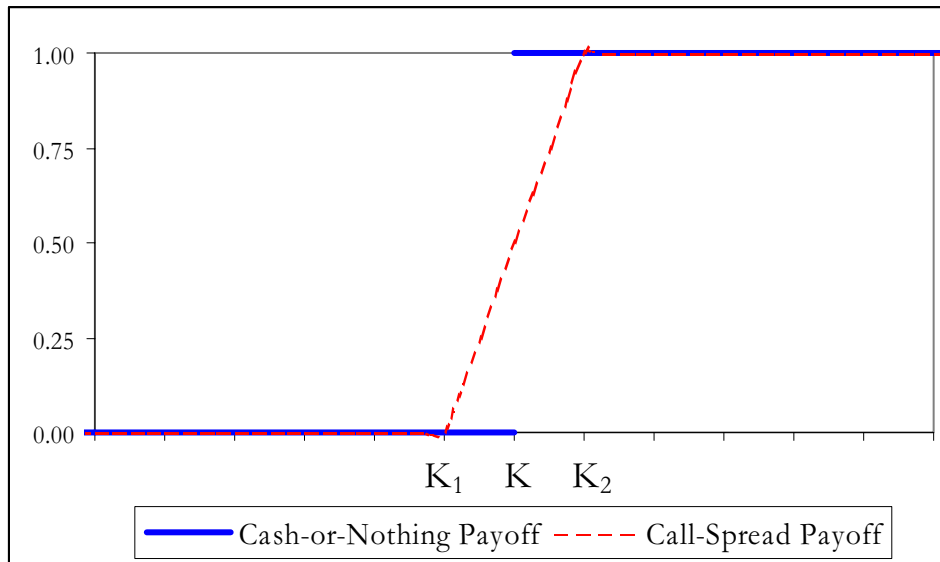
One method that can be used to hedge a cash-or-nothing option is to use a spread of standard options. For a cash-or-nothing call with strike price K , consider the following strategy:

- a long call, strike $K_1 < K$;
- a short call, strike $K_2 > K$.

If we hold a face value of $1/(K_2 - K_1)$ of each of these options, the payoff for the strategy is

$$V(T) = \begin{cases} 1 & \text{if } S(T) > K_2 \\ \frac{S(T) - K_1}{K_2 - K_1} & \text{if } K_1 < S(T) \leq K_2 \\ 0 & \text{if } S(T) \leq K_1. \end{cases}$$

Figure I.B.9.4: Call Spread Digital Hedge



The further K_1 and K_2 are from K , the greater the basis risk that remains between the cash-or-nothing payoff and the hedge. As K_1 and K_2 get closer to K , the payoff of the call spread approaches that of the cash-or-nothing call. However, as K_1 goes to K_2 the face value $1/(K_2 - K_1)$ becomes huge; the associated transaction costs could be prohibitive.

There is really no effective way of neutralising the large gamma at the strike price associated with these options. The same problem arises with any option in which a digital option is incorporated, such as the contingent premium options above (see Section I.B.9.5.4). In general, any discontinuous payoff, such as a barrier option that is in-the-money when the barrier is hit (see Section I.B.9.12.1), is similarly difficult to manage.

I.B.9.6 Two Asset Options

There are many exotics that depend on two different assets. The term *two-colour rainbow* option was first used by Mark Rubinstein (1991) to describe options involving two correlated asset prices, which cannot be valued as if they were a single asset. In this section we look at some simple path-independent cases of these.

I.B.9.6.1 Product and Quotient Options

The payoff of a *product option* at expiry T is

$$V(T) = [\lambda(S_1(T)S_2(T) - K)]^+.$$

Because asset prices in the BSM model are lognormally distributed and the product of two jointly lognormal variables is also lognormal, the BSM formula can be applied to the product $S_1(T)S_2(T)$. Although the payoff on the product of two different prices may not have much commercial appeal, this payoff can be useful in deriving formulae for some other types of options.

A similar payoff, which makes more business sense, is a *quotient option*, whose payoff is

$$V(T) = [\lambda(S_1(T)/S_2(T) - K)]^+.$$

For instance, a *share ratio option* is designed to capture the relative performance of an individual stock price $S_1(T)$ compared to a stock index $S_2(T)$. The strike price is chosen to be the ratio between the stock and the index at the inception of the contract, $K = S_1(0)/S_2(0)$.

I.B.9.6.2 Exchange Options

These give the holder the right to exchange an amount of one asset for an amount of another asset at expiry, if it is advantageous to do so. That is, the payoff at expiry is

$$V(T) = [n_1S_1(T) - n_2S_2(T)]^+,$$

where n_1 is the number of units of the first asset and n_2 is the number of units of the second asset specified in the contract. Exchange options also have an analytical formula within the BSM model (Margrabe, 1978).

I.B.9.6.3 Outperformance Options

Quotient options and exchange options both give the holder a payoff if the first asset performs better than the second. Options which give the holder the ‘better’ of two assets in some sense are called *outperformance options*. Other ways of achieving this include:

Option to Obtain the Maximum of Two Assets:

$$V(T) = \max[S_1(T), S_2(T)].$$

Notice that this payoff can be rewritten as

$$V(T) = S_1(T) + \max[S_2(T) - S_1(T), 0].$$

So this option is just a portfolio containing the underlying asset and an exchange option.

Relative Performance Option:

$$V(T) = [S_1(T)/S_1(0) - S_2(T)/S_2(0) - K]^+.$$

These can be used by fund managers to ensure they achieve a benchmark return, while taking a view that a particular instrument will outperform the benchmark. For instance, the fund needs to match an index $S_2(T)$ but the manager believes a particular stock $S_1(T)$ will outperform the index in the period. By buying the relative performance option, he can leave the structure of his portfolio otherwise unchanged, benefit if his view is correct, but still match the benchmark return (less the outlay of the option premium) if it is not. Relative performance options are a special case of spread options (Section I.B.9.6.5).

I.B.9.6.4 Other Two-Colour Rainbow Options

Payoffs include:

Option Delivering Better of Two Assets and Cash:

$$V(T) = \max[S_1(T), S_2(T), K].$$

Call on Maximum of Two Risky Assets:

$$\begin{aligned} V(T) &= [\max[S_1(T), S_2(T)] - K]^+ \\ &= \max[S_1(T), S_2(T), K] - K. \end{aligned}$$

Put on Maximum of Two Risky Assets:

$$\begin{aligned} V(T) &= [K - \max[S_1(T), S_2(T)]]^+ \\ &= \max[S_1(T), S_2(T), K] - \max[S_1(T), S_2(T)]. \end{aligned}$$

There are analytical pricing formulae for each of these options in the BSM model that involve the bivariate normal distribution (see Chapter II.E). There are also formulae for quotient and exchange options. However, as we shall see for spread options in the next section, this is not always the case.

I.B.9.6.5 Spread Options

A spread option is an option on the spread or difference between two prices or rates. They are a particularly important variety of two-asset option because they are so widespread – for instance, options are traded on the difference between:

- the return on an individual stock and on a stock index;

- interest rates in the same economy or different economies;
- futures contracts with different maturities (the *calendar* spread);
- raw and refined grades of a commodity, such as
 - the *crack* spread in the oil markets,
 - the *crush* spread in soft commodities,
 - the *spark* spread in the electricity markets.

The general payoff at expiry for an option on the spread between two asset prices is

$$V(T) = [\lambda(\alpha_1 S_1(T) - \alpha_2 S_2(T) - K)]^+.$$

Here the scaling factors α_1 and α_2 are both positive; as usual, $\lambda = 1$ for a call and $\lambda = -1$ for a put. The strike price K refers to a possible future value of the spread.

Although this payoff looks no more complicated than the rainbow payoffs above, there is no analytical pricing formula in the BSM model, because the difference of two lognormal variables does not follow the lognormal or any other well-known distribution.

I.B.9.6.6 Correlation Risk

The *vega* is the sensitivity of an option price to volatility (see Chapter I.A.5). Each of the two asset options discussed above has two vegas, one with respect to each asset price volatility, σ_{S_i} :

$$\partial V / \partial \sigma_{S_i}.$$

The sensitivity to the correlation ρ_{S_1, S_2} between the log-asset returns is⁷

$$\partial V / \partial \rho_{S_1, S_2}.$$

These sensitivities can only be hedged using another instrument with a like sensitivity. For example, $\partial V / \partial \sigma_{S_i}$ can be hedged using a vanilla option on S_i . We need to hold a face value of the vanilla option given by the ratio of the vegas:

$$\frac{\partial V_{\text{rainbow}} / \partial \sigma_{S_i}}{\partial V_{\text{vanilla}} / \partial \sigma_{S_i}}.$$

Likewise, the correlation sensitivity can only be hedged by taking a position in another correlation product, such as another two-asset option, of face value

$$\frac{\partial V_{\text{rainbow}} / \partial \rho_{S_1, S_2}}{\partial V_{\text{hedge}} / \partial \rho_{S_1, S_2}}.$$

⁷ There does not seem to be a universally accepted term for the sensitivity to correlation.

The practical problem is finding another two-asset option in the market to use as a hedge.

I.B.9.7 Quantos

This section deals with two-asset options in which one asset is in a foreign economy and its price is in the foreign currency; the other asset is the exchange rate between the domestic and foreign currencies. We change the usual notation slightly here to cater for this:

- S_t is the price *in foreign currency* of a foreign asset at time t , for example a European stock in euro (EUR) when we are accounting in US dollars (USD).
- X_t is the exchange rate at time t , that is, the price of one unit of the foreign currency in domestic currency, for example the price of 1 EUR in USD.

Sometimes the guaranteed exchange rate options in Section I.B.9.7.5 are referred to as ‘quantos’; sometimes the term ‘quanto’ is used for any such currency-foreign-asset hybrid. We use ‘quanto’ in this more general sense: the face value or ‘quantity’ of the instrument may vary, depending on the price of a different asset.

I.B.9.7.1 Foreign Asset Option Struck in Foreign Currency

An investor may want to protect himself against changes in the price of a foreign asset. He could do this by simply buying a vanilla option on the foreign asset, in the foreign economy. Its payoff *in foreign currency* is, as usual,

$$V_S(T) = [\lambda(S_T - K_S)]^+.$$

However, the investor is accounting in domestic currency so the payoff needs to be translated back into domestic currency:

$$V(T) = X_T [\lambda(S_T - K_S)]^+.$$

The strike price K_S is in foreign currency per unit of foreign asset. The price of the option now in domestic currency is just the usual price of a vanilla option in the foreign economy, translated into domestic currency at the current exchange rate. For example, buying a vanilla option on the Nikkei in Japan and translating its value into USD at the current rate.

I.B.9.7.2 Foreign Asset Option Struck in Domestic Currency

The investor may want protection against price changes in the foreign asset but wants to know his exposure in local currency terms. Setting the strike of the option in domestic currency allows this. The option payoff is

$$V(T) = [\lambda(X_T S_T - K_{XS})]^+.$$

Here both $X_T S_T$ and the strike price K_{XS} are in terms of domestic currency per unit of foreign asset. This payoff is very similar to that of the product option in Section I.B.9.6.1, except that here S_T is in foreign currency.

For example, an Australian gold miner is exposed to the international price of gold in USD but ultimately accounts in Australian dollars (AUD). He may prefer an option whose strike price is set in 1 oz gold in AUD, that is, an option on the product:

$$(1 \text{ oz Gold} / \text{AUD}) = (1 \text{ oz Gold} / \text{USD}) \times (1 \text{ USD} / \text{AUD}).$$

I.B.9.7.3 Implied Correlation

Cross-currency rates are exchange rates in which neither currency is the USD, for example, the EUR/CAD (euro–Canadian dollar rate). A common use of ‘foreign asset options struck in foreign currency’ is the management of cross-currency options. The currency pairs EUR/USD and CAD/USD are more liquid than the cross EUR/CAD and their implied volatilities are more easily observed. Since

$$(1 \text{ EUR} / \text{CAD}) = (1 \text{ EUR} / \text{USD}) \times (1 \text{ USD} / \text{CAD}),$$

a ‘standard’ foreign exchange option on the cross-currency pair can be managed as an option on the product of the two more liquid currency pairs, with volatility given by

$$\begin{aligned}\sigma_{XS}^2 &= \sigma_X^2 + \sigma_S^2 + 2 \rho_{X,S} \sigma_X \sigma_S, \\ \sigma_{XS} &= \sigma_{\text{EUR/CAD}}, \\ \sigma_S &= \sigma_{\text{EUR/USD}}, \\ \sigma_X &= \sigma_{\text{USD/CAD}}, \\ \rho_{X,S} &= \rho_{\text{EUR/USD, USD/CAD}}.\end{aligned}$$

It may be possible to observe implied volatilities for $\sigma_{\text{EUR/CAD}}$ as well as $\sigma_{\text{USD/CAD}}$ and $\sigma_{\text{EUR/USD}}$ from vanilla options. We could then rearrange the equation to get the ‘implied correlation’:

$$\rho_{X,S} = \frac{\sigma_{XS}^2 - \sigma_X^2 - \sigma_S^2}{2 \sigma_X \sigma_S}.$$

This also implies the correlation risk could be hedged by hedging the vegas associated with σ_X , σ_S and σ_{XS} , if liquid vanilla options are available in each currency pair.

I.B.9.7.4 Foreign Asset Linked Currency Option

An international fund manager may want exposure to the prices of foreign stocks to diversify his portfolio, but may also want to hedge the exchange rate at which he will later repatriate his investment. The amount of foreign currency he will need will depend on the foreign asset price. This leads to the option payoff

$$V(T) = S_T[\lambda(X_T - K_X)]^+.$$

This is the payoff of a foreign exchange option whose face value is the future foreign asset price, so the strike price K_X is in domestic currency per unit of foreign currency.

I.B.9.7.5 Guaranteed Exchange Rate Foreign Asset Options

An investor may want to protect his investment in a foreign market with a foreign asset option. To avoid exposure to the exchange rate, he could buy a guaranteed exchange rate foreign asset option. This incorporates conversion of the payoff from the foreign asset option into domestic currency at a pre-agreed fixed rate, say G_X . Because of the fixed exchange rate, these options are also known as *currency immunized options* and sometimes just *quantos*. Their payoff is

$$V(T) = G_X[\lambda(S_T - K_S)]^+,$$

where G_X is an exchange rate, in domestic currency per unit of foreign currency.

Sometimes the ‘guaranteed exchange rate’ does not really look like an exchange rate. For instance, Nikkei puts have been listed on the American Stock Exchange which give the payoff of a put on the Nikkei, paid in USD. The Chicago Mercantile Exchange has a similar futures contract; for S_T the Nikkei Index at maturity and K_S the futures contract rate, its payoff is

$$\lambda(S_T - K_S) \times 5 \text{ USD}.$$

I.B.9.8 Second-Order Contracts

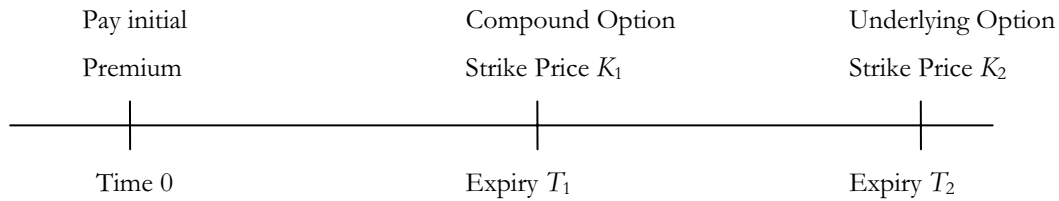
In a second-order contract, the underlying asset is another option.

I.B.9.8.1 Compound Options

Compound options are the simplest second-order contracts, being a vanilla option on an underlying vanilla option; that is, they may be:

- call on a call,
- call on a put,
- put on a call,
- put on a put.

Suppose the compound option has strike price K_1 and expires at T_1 . It is an option to buy/sell the underlying option; the underlying option expires at $T_2 > T_1$ and has strike price K_2 . At T_1 , the underlying option has $T_2 - T_1$ left to run.



If we write $V^c(t, S(t), K, \tau)$ and $V^p(t, S(t), K, \tau)$ for the prices at time t of a standard call and put with strike price K and time τ until expiry, then the compound option payoffs at its expiry T_1 are:

Call-on-a-Call: $[V^c(T_1, S(T_1), K_2, T_2 - T_1) - K_1]^+$,

Call-on-a-Put: $[V^p(T_1, S(T_1), K_2, T_2 - T_1) - K_1]^+$,

Put-on-a-Call: $[K_1 - V^c(T_1, S(T_1), K_2, T_2 - T_1)]^+$,

Put-on-a-Put: $[K_1 - V^p(T_1, S(T_1), K_2, T_2 - T_1)]^+$.

The premium on the compound option is paid up-front at time 0. It is cheaper than the premium that would be paid at time 0 for a standard option with strike K_2 expiring at T_2 . The strike on the compound option is a premium for the underlying option, to be paid at T_1 . If the compound option is exercised and both premiums are paid, the holder ends up paying more than the price of the equivalent standard option, as you would expect from having the additional right to let the compound option expire without exercise. Compound options do not seem to be very popular in practice, possibly because they are more expensive than the equivalent standard option by the time both premiums are paid.

I.B.9.8.2 Typical Uses of Compound Options

A standard use of a compound option is when a company is involved in a tender. The company may want protection against adverse price movements if the tender is successful, but not otherwise. For example, an exporter tendering for a contract in foreign currency may want protection against the foreign currency weakening; he can buy a compound call, to buy a put on the exchange rate. The compound call expires when the result of the tender is known; the underlying put expires when the exporter would receive payment for his goods in foreign currency. He is unwilling to pay a large up-front premium for protection he may not ultimately need. If the exporter wins the tender, he can exercise the compound option and buy the foreign exchange put for the agreed premium (the compound strike) or buy the same put in the market if it is cheaper (the compound option expires out-of-the-money). If the exporter does not win the tender, he could let the compound option lapse, or if it expires in-the-money he would exercise it, buying the underlying put for the compound strike and then selling it in the market where it is worth more.

Compound options are also a way of taking a view on future levels of volatility. Suppose a hedger requires a standard option but believes that current volatilities are too high and will not persist. If his view is correct, it will be cheaper to buy the option later. The compound option costs less than buying the underlying option outright. If his view is correct, the compound option will expire out-of-the-money and he will be able to buy the standard option cheaper in the market, but if he is wrong he has locked in a premium on the option he needs at current volatilities.

The underlying option could also be an exotic option, such as a call on an average rate option, but this seems to be unusual in practice.

I.B.9.8.3 Instalment Options

These allow the holder to pay the option premium in a number of instalments. At any instalment date, the holder may choose to:

- make a payment to continue with the option, or
- not pay the instalment and let the option lapse.

Instalment options allow the cost of protection to be spread over a series of dates. The initial upfront premium is lower than the price of the equivalent standard option, but the total cost of the present value of all instalments will be more. They have the advantage that if the market moves significantly away from the level at which protection has been set, the holder can let the option lapse and enter a new hedge arrangement.

Compound options can be thought of as a special case of instalment options, in which there is a single instalment, namely the compound option's strike. Likewise, instalment options are a nested series of options on options; at each date, the holder can exercise the option to 'buy' the remainder of the series.

As with other options where the holder may exercise a choice (see Section I.B.9.9), they can best be priced using a tree or lattice method.

I.B.9.8.4 Extendible Options

There are two main varieties of extendible options. Both allow either the holder or the writer of the option to extend the option at an original expiry date T_1 until a later expiry date T_2 .

In a *holder-extendible* option, the holder may choose to pay an additional premium \mathcal{A} at T_1 to extend the expiry from T_1 until T_2 . Using the same notation as in Section I.B.9.8.1 for compound options, its payoff at T_1 is

$$\max [S(T_1) - K, V^c(T_1, S(T_1), K, T_2 - T_1) - \mathcal{A}, 0]$$

for a call, and

$$\max [K - S(T_1), V^p(T_1, S(T_1), K, T_2 - T_1) - \mathcal{A}, 0].$$

for a put. Comparing these to the compound payoffs,

$$\max [V^c(T_1, S(T_1), K_2, T_2 - T_1) - K_1, 0]$$

for a call on a call, and

$$\max [V^p(T_1, S(T_1), K_2, T_2 - T_1) - K_1, 0],$$

for a call on a put, we can see that with strike $K_1 = \mathcal{A}$, the payoff on the holder-extendible option is potentially more, so its value will be more than the value of the equivalent ‘call-on’-compound option.

In a *writer-extendible* option, at T_1 , if the option is in-the-money, it is exercised in the usual way; if the option is out-of-the-money, the expiry is automatically extended until T_2 . The payoff at T_1 for a call is

$$V(t) = \begin{cases} S(T_1) - K & S(T_1) \geq K \\ V^c(T_1, S(T_1), K, T_2 - T_1) & \text{otherwise;} \end{cases}$$

while for a put it is

$$V(t) = \begin{cases} K - S(T_1) & S(T_1) \leq K \\ V^p(T_1, S(T_1), K, T_2 - T_1) & \text{otherwise.} \end{cases}$$

The extension is like a compensation for the holder for the option ending up out-of-the-money. So this contract must be worth more than the equivalent standard option.

Within the BSM model, both types of extendible options have analytical formulae. However, like compound and instalment options, they really depend on future levels of volatility, so may be better priced using a model that takes account of future volatility uncertainty.

I.B.9.9 Decision Options

The options in this section give the holder the right to make a decision affecting the payoff of the option or its timing. The easiest way to price such options is to use a tree or lattice method, because we work backwards in time through the tree and so can determine the best time to make the decision.

I.B.9.9.1 American Options

A *European* option can only be exercised at expiry. In contrast, the holder of an *American* option can exercise it at any time during its life. Consequently it must always be worth more than its payoff if it were exercised, that is, its intrinsic value.

The binomial tree model was originally developed to value the extra premium paid for the right to exercise early. At each node in the tree, the option price is increased if necessary to be at least its exercise value.

I.B.9.9.2 Bermudan Options

These are also called *mid-Atlantic* or *limited exercise* options. As the name suggests, they are ‘between’ a European and an American option: they can be exercised early, but only at certain times, usually a discrete set of dates or, more generally, during specified periods of time. They can also be priced using a tree model, but the adjustment to increase the option price to the exercise value is only made at nodes where an exercise date falls during the subsequent time interval.

Because it gives the holder more, the equivalent American option is always worth more than the Bermudan option; the equivalent European option is always worth less. The greater the number of exercise dates, the closer it will be to the American option price; the fewer the number of exercise dates, the closer it will be to the European option price.

Options given to staff by a company are often Bermudan. Call features attached to corporate bonds are sometimes Bermudan.

I.B.9.9.3 Shout Options

The holder of a shout option has the extra right, which may be exercised once only, to reset the strike price to be at-the-money. When the holder of the option exercises his right to reset the strike, the option becomes a standard option. Consequently, it must always be worth more than the equivalent vanilla option. Shout options are also most easily priced using a binomial tree algorithm: at each node, the option price is adjusted to be at least the value of the equivalent vanilla option at that point. Variations can occur in practice, for example:

- the period during which the holder of the option may set the strike price may be restricted, known as ‘shouting through a window’;
- the strike may be reset to a proportion α of the current asset price;
- more general multiple-shout options, in which the holder may choose when to adjust the strike more than once, are also possible.

I.B.9.10 Average Options

Average options have path-dependent payoffs based on averages of asset prices observed during the life of the option.

I.B.9.10.1 Average Rate and Average Strike Options

The most common type are *average rate options*, with payoff at expiry

$$V(T) = [\lambda(A - K)]^+.$$

Less common are *average strike options*, with payoff at expiry

$$V(T) = [\lambda(S_T - A)]^+.$$

Here A is an average of asset prices taken over some part of the life of the option and K is the strike. These are like the standard option payoffs, with the average replacing the final asset price in the average rate payoff and the average replacing the strike in the average strike payoff.

Usually the average is the *arithmetic average* of asset prices at some pre-agreed times T_1, \dots, T_n :

$$A_n = \sum_{i=1}^n w_i S(T_i)$$

where the weights w_1, \dots, w_n are positive and sum to 1.

Arithmetic average options are also called *Asian* options, although sometimes this term is reserved for the special case of equal weights, that is:

$$A_n = \frac{1}{n} \sum_{i=1}^n S(T_i).$$

An average rate option with unequal weights in the average is sometimes referred to as a *flexible* Asian or average rate option. Less commonly, the average is the *geometric average* of the prices:

$$G_n = \prod_{i=1}^n (S(T_i))^{w_i}.$$

Currency indices such as the New Zealand Trade Weighted Index often use a geometric average.

I.B.9.10.2 Motivations and Uses

Average rate options are typically used by a company that has periodic payments or receivables, and wants to gain cost-effective trend protection against unfavourable asset price movements. The average can be tailored to match a company's cash flows. They are particularly common in the markets for oils and other commodities.

Their appeal lies in the fact that average rate options are less expensive than a strip of vanilla options with the same strike price, expiring at each of the averaging dates, because some of the individual options may give a payoff when the option on the average does not:

$$\max \left[\lambda \left(\sum_{i=1}^n w_i S(T_i) - K \right), 0 \right] \leq \sum_{i=1}^n w_i \max \left[\lambda (S(T_i) - K), 0 \right].$$

Also the average asset price is less volatile than the underlying asset price, so an average price option will usually be cheaper than a single standard option with the same strike and expiry, but this also depends on the shape of the forward asset price curve.

An average rate option would be used by a company with an invoiced cost for materials based on the average of asset prices over a definite period, for example oil consumers. Average strike options are less common but could be used by a company with a foreign parent which is required to translate profits to the parent at the average exchange rate on set balance sheet dates.

I.B.9.10.3 Other Options Involving Averages

Double average rate options

These have pay-off

$$V(T) = [\lambda(\mathcal{A}_2 - \mathcal{A}_1)]^+.$$

Here \mathcal{A}_1 is an average over a near-dated period and \mathcal{A}_2 is an average over a far-dated period, where the periods do not overlap. Notice that average strike options are a special case of double average rate options with $\mathcal{A}_2 = S(T)$. Also, once the near period has passed \mathcal{A}_1 is fixed, so they become average rate options. These are used by large multinational corporations, such as McDonald's and Microsoft, to hedge repatriation of their foreign subsidiaries' profits in one quarter against the same quarter in the previous calendar year.

Hawaiian options

This is the name given to *American Asian* options. If exercised early, the payoff is based on the average so far; for instance, the arithmetic average so far at time t is

$$A_{n_t} = \frac{\sum_{i=1}^{n_t} w_i S(T_i)}{\sum_{i=1}^{n_t} w_i},$$

where n_t is the number of averaging dates that have passed up to and including time t . For equal weights, this becomes

$$A_{n_t} = \frac{1}{n_t} \sum_{i=1}^{n_t} S(T_i).$$

'Australian' Asian options

This name has been used for options on the ratio of the average to the final price, traded at the Australian Stock Exchange since 1992.

Inverse averages

Sometimes in the foreign exchange markets, a counterparty requires an inverse average payoff:

$$V(T) = \left[\lambda \left(\frac{1}{A} - \frac{1}{K} \right) \right]^+$$

This arises because of foreign exchange quotation conventions. For example, USD–JPY rates are quoted in JPY per USD. Typically, a US company is hedging a JPY face value and wants to pay for the option in USD, but wants a payoff based on an arithmetic average of market quotes; the average asset price A and the strike K are in JPY per USD.

I.B.9.10.4 Pricing and Hedging Average Options

As for spread options, the distribution of the arithmetic average of asset prices in the BSM model is not known and there is consequently no simple analytical formula for the current price of an arithmetic average option. However, the distribution of the geometric average is lognormal and there is a formula for the price of a geometric average option. Average rate options can be priced efficiently using Monte Carlo simulation; for arithmetic average options, the equivalent geometric average payoff is used to improve the accuracy of the simulation. A number of accurate analytical approximations are also available.

American exercise makes pricing more difficult, because it is awkward to handle the early exercise feature using simulation, but to value a path-dependent option such as an average rate option using a binomial tree requires a method of avoiding evaluation along every path through the tree.

During its life, the delta of an average rate option will make a discrete fall each time an averaging date is reached. Each time part of the average is set, more of the payoff is determined and less risk remains. When almost all of the average has been set, changes in the asset price will have little effect on the option's value and the delta will approach zero. Conversely, average strike options have very little delta before the averaging begins. Once the average has been set for an average strike option, it becomes a standard option. The delta ratchets towards a standard option delta each time an averaging date is reached and part of the average is set.

I.B.9.11 Options on Baskets of Assets

I.B.9.11.1 Basket Options

A *basket option* is a standard option on a portfolio or basket of assets. Suppose we have N component assets whose prices at time t are $S_1(t), \dots, S_N(t)$, and n_i is the face value or amount of the i th asset. The price of the basket is

$$B(t) = \sum_{i=1}^N n_i S_i(t).$$

The payoff of the basket option at expiry T is

$$V(T) = [\lambda(B(T) - K)]^+.$$

Basket options can be a cost-effective way of protecting against exposure to the prices of a number of different assets with a single instrument. As long as the assets are not perfectly correlated, the basket option will cost less than a portfolio of individual vanilla options on each asset that are the same proportion in- or out-of-the-money. The lower the correlation between the individual assets, the greater the cost saving.

I.B.9.11.2 Pricing and Hedging Basket Options

Like arithmetic average options which are also based on a sum of asset prices, there is no simple analytical formula for pricing basket options in the BSM model, but they can be priced efficiently using Monte Carlo simulation provided the number of assets is not too large.

The basket option is a multi-asset option, so to hedge the basket option we need to hold a portfolio of the N underlying assets. In the BSM model (and other complete market models) we know from Chapter I.A.8 that each delta is given by the derivative of the option price with respect to the i th asset price:

$$\Delta_i = \frac{\partial V}{\partial S_i}.$$

The i th asset price *gamma* is the change in the i th delta as the i th underlying asset price changes:

$$\Gamma_i = \frac{\partial^2 V}{\partial (S_i)^2}.$$

For basket options, because the assets prices are correlated, every delta will change when each asset price changes. This gives rise to the *cross gamma*, the change in the j th delta as the i th asset price changes:

$$\Gamma_{i,j} = \frac{\partial^2 V}{\partial S_i \partial S_j}.$$

Because the real world is *not* the perfect BSM world, but does have transaction costs and jumps in the asset prices, the fact that every delta changes when each asset price changes means that basket options (and multi-asset options in general) can be quite difficult to manage in practice.

The exposure to the correlation estimate can only be hedged using another instrument that also depends on the correlation. Likewise, to hedge the cross-gamma risk requires another instrument that is sensitive to two different asset prices, with its own cross gamma. So it is not possible for the seller of a basket option to hedge these risks using only vanilla single-asset options. In some markets, such as the foreign exchange markets, finding another two-asset option is relatively easy, for instance a cross-currency option (Section I.B.9.7.3). But this may be more difficult to achieve in other markets such as the commodity or equity markets. For an optimal hedge using only single-asset options, see Ashraff *et al.* (1995).

I.B.9.11.3 Mountain Options

These are complex path-dependent options based on baskets of stocks that have recently become very popular in the structured equities markets. They are marketed under the names of various mountains. For instance, an *Everest* option gives the holder a payoff based on the worst-performing member of a large basket of stocks (typically 10–25 stocks) with long expiry (possibly 10–15 years). An *Atlas* option is a call on a basket of stocks where the m_1 worst performing stocks and the m_2 best performing stocks are excluded from the payoff. There are many others in the ‘range’ including *Altiplano*, *Annapurna* and *Himalayan* options.

Because of their path-dependency, they are best priced using simulation. Their multi-asset payoffs make them particularly sensitive to assumptions about correlation between the stocks. For more details see Lipton (2003, Chapters 28–29).

I.B.9.12 Barrier and Related Options

Barrier options are one of the most popular types of exotic, particularly in the foreign exchange markets. This popularity is partly due to the fact that barrier options are always cheaper than the equivalent vanilla option. The number and style of the barriers can be tailored to suit the needs of a hedger or the views of a speculator. The payoff of a barrier option depends on whether the asset price reaches a certain level, called the *barrier* or *trigger*, during a given time period.

I.B.9.12.1 Single-Barrier Options

These are options with a single barrier which may be either *knock-in* or *knock-out*. If the asset price reaches a knock-out barrier, the option ceases to exist. When the asset price reaches a knock-in barrier, the option becomes a vanilla option; a payoff will subsequently occur at expiry if the

usual exercise conditions are met. The vanilla option that knocks in or out can be a call or a put. If the barrier is above the current asset price, it is referred to as an *up* barrier; if it is below the current asset price, it is a *down* barrier. For pricing and hedging it also matters whether the option is in-the-money or out-of-the-money relative to the strike price when the barrier is hit. A barrier option that is in-the-money when the barrier is hit is referred to as *reverse* knock-in or knock-out.

This gives $2 \times 2 \times 2 \times 2 = 16$ possible types of single barrier options, but actually there are only 12 since:

- an up-and-out out-of-the-money call always gives a zero payoff;
- an up-and-in out-of-the-money call gives the same payoff as a standard call;
- similarly, a down-and-out out-of-the-money put always gives a zero payoff;
- a down-and-in out-of-the-money put gives the same payoff as a standard put.

The payoff for each single-barrier option may be written in the form:

$$V(T) = \mathbf{I}_{\mathcal{A}} [\lambda(S(T) - K)]^+.$$

\mathcal{A} is the condition that the barrier is hit or not:

- Down-and-in: $\min_{\{0 \leq t \leq T\}} S(t) \leq L$
- Down-and-out: $\min_{\{0 \leq t \leq T\}} S(t) > L$
- Up-and-in: $\max_{\{0 \leq t \leq T\}} S(t) \geq H$
- Up-and-out: $\max_{\{0 \leq t \leq T\}} S(t) < H$

The down barrier L is below the current asset price $S(0)$; the up barrier H is above $S(0)$.

Within the BSM model, there is an analytical formula for the price of each single-barrier option. This is another reason why they were initially quite popular in practice. While they are relatively easy to price in the BSM model, they can be difficult to manage. The delta and gamma can move rapidly and/or change sign as the asset price approaches and goes through the barrier. Unlike vanilla options, the delta of a barrier option can be greater than 1 (or less than -1); that is, the face value of the required hedge may be larger, possibly by several times, than the contract face value. The delta can be discontinuous at the barrier, so a large hedge position may need to be instantaneously unwound when the barrier is hit. If the option is in-the-money when the barrier is hit, it instantaneously acquires or loses a positive amount of intrinsic value. The value of such an option behaves as if it contains a digital component, and the provider of the option can experience similar difficulties to those associated with managing a cash-or-nothing option. For example, the charts below show the BSM model value and delta for an up-and-out call with strike 100, barrier 120 and 1 week to expiry, as the asset price goes through the barrier.

Figure I.B.9.5: Up-and-Out Call Value

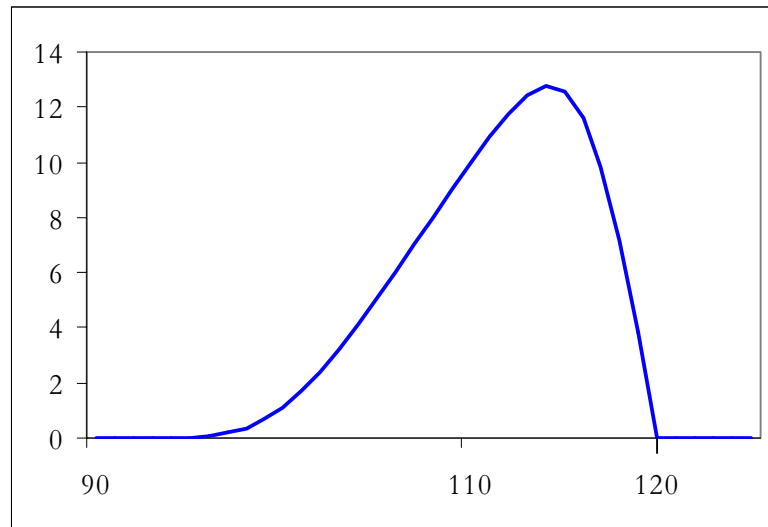
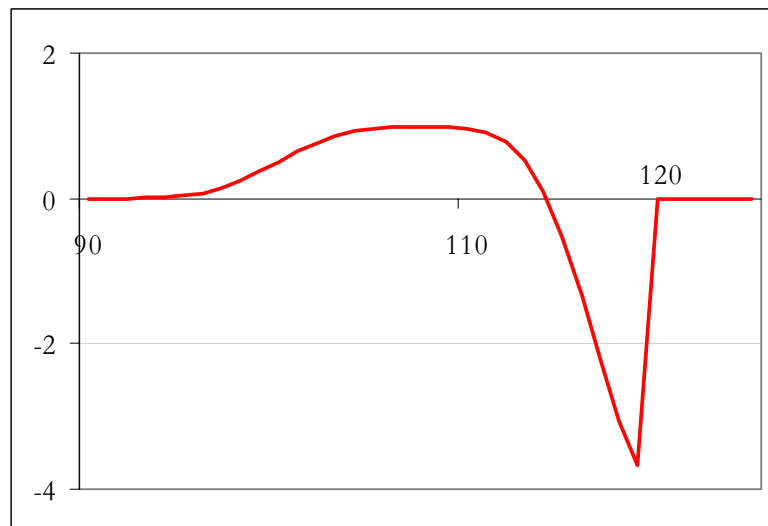


Figure I.B.9.6: Up-and-Out Call Delta



I.B.9.12.2 No-Touch, One-Touch and Rebates

A *no-touch* option pays a fixed amount at expiry if the barrier is *not* hit. A *one-touch* option pays a fixed amount if the barrier *is* hit; the payment may occur either at the option expiry or when the barrier is hit. For a fixed amount α the payoff is just

$$V(\tau) = \alpha \mathbf{I}_{\mathcal{A}}.$$

The barrier conditions \mathcal{A} are the same as above for a single-barrier option, and τ is either the hitting time or the expiry T , depending on when the payoff occurs.

A *rebate* is a cash payment to compensate the holder when the option does not give a payoff because the barrier condition has not been met. Knock-in options may pay a rebate at expiry if

the barrier is *not* hit; knock-out options may pay a rebate if the barrier *is* hit. They may be viewed as compensation for no payoff to the holder if the option does not knock in, or to offset a new premium if an option intended as a hedge has knocked out. In effect a rebate is a no-touch or one-touch option attached to the primary barrier option, so the holder does pay for any ‘compensation’ received in the form of a rebate by paying up-front for this additional option.

I.B.9.12.3 Partial-Barrier Options

In the single-barrier payoffs described above, hitting the barrier may take effect at any instant during the life of the option. More generally, a barrier may have a period of time during which the barrier is *active*, that is, if the barrier is hit during this period then the option will be knocked in or knocked out, but outside this period hitting the barrier has no effect; the barrier is then said to be *protected*. The barrier is said to be a *full-life* barrier if it is active for the whole period from the date at which we are valuing the option until the expiry date of the option; otherwise it is termed a *partial* barrier.

Writing time T_1 for the start of the barrier period and time T_2 for the end of the barrier period, payoffs for single partial barriers are also given by

$$V(T) = \mathbf{I}_{\mathcal{A}} [\lambda(S(T) - K)]^+.$$

Only the barrier conditions \mathcal{A} change:

- Down-and-in: $\min_{\{T_1 \leq t \leq T_2\}} S(t) \leq L$
- Down-and-out: $\min_{\{T_1 \leq t \leq T_2\}} S(t) > L$
- Up-and-in: $\max_{\{T_1 \leq t \leq T_2\}} S(t) \geq H$
- Up-and-out: $\max_{\{T_1 \leq t \leq T_2\}} S(t) < H$

A *start-of-period* barrier is active from today until some future date before the expiry date: $T_1 = 0$ and $T_2 < T$. After this date, a knock-out option becomes a standard option and a knock-in option ceases to exist. For instance, the holder may not want his hedge to be knocked out as the expiry date approaches.

An *end-of-period* partial barrier is protected from today until some future date and then active from that date until expiry: $T_1 > 0$ and $T_2 = T$. A motivation for an end-of-period barrier could be to protect the holder from expected short-lived volatility, for example as could be caused by an upcoming election. Start-of-period and end-of-period barriers are the most common partial barriers. They have analytical formulae in the BSM model in terms of the bivariate normal distribution.

I.B.9.12.4 Double-Barrier Options

Double-barrier options have both an upper barrier H above the current asset price and a lower barrier L below it. They may be:

- *double knock-out*: both barriers must not be hit;
- *double knock-in*: at least one barrier must be hit;
- *up-and-out, down-and-in*: an up-and-out option is created if the down barrier is hit, but the option will cease to exist if the up barrier is hit;
- *up-and-in, down-and-out*: a down-and-out option is created if the up barrier is hit, but the option will cease to exist if the down barrier is hit.

Payoffs for double-barrier options can again be written as

$$V(T) = \mathbf{I}_A [\lambda(S(T) - K)]^+.$$

Now the barrier conditions become:

- Double-out: $L < \min_{\{0 \leq t \leq T\}} S(t) \leq \max_{\{0 \leq t \leq T\}} S(t) < H$
- Double-in: $\min_{\{0 \leq t \leq T\}} S(t) \leq L$ or $\max_{\{0 \leq t \leq T\}} S(t) \geq H$
- Up-and-out, down-and-in: $\min_{\{0 \leq t \leq T\}} S(t) \leq L$ and $\max_{\{0 \leq t \leq T\}} S(t) < H$
- Up-and-in, down-and-out: $\min_{\{0 \leq t \leq T\}} S(t) > L$ and $\max_{\{0 \leq t \leq T\}} S(t) \geq H$

I.B.9.12.5 Even More Barrier Options

When the barrier is active for a period in the middle of the option's life, $0 < T_1 \leq T_2 < T$, it is called a *window barrier*. In the case of double-barrier options, there may be different active-barrier periods for the up barrier and the down barrier.

In the payoffs above, the barrier is *continuous*: it may be triggered at any instant. *Discrete-barrier options* have an active-barrier period consisting of a discrete set of dates. The active-barrier period could be a series of windows and the barrier could be different for each period, but this is less common.

A *two-asset* or *outside* barrier option is a correlation product in which the barrier refers to one asset price while the option payoff refers to another asset price:

$$V(T) = \mathbf{I}_A [\lambda(S_1(T) - K)]^+.$$

The barrier condition depends on a second asset price $S_2(t)$, for instance, for a single up-and-out outside barrier the condition is: $\max_{\{0 \leq t \leq T\}} S_2(t) < H$.

Outside-barrier options with a currency component particularly make business sense, for example

a currency option that will knock out if the price of a commodity reaches a given level. For instance, an Australian gold miner sells his product into the market in USD so may require protection against the AUD–USD rate; however, below a certain gold price the mine may not be viable so no currency protection would be necessary. Alternatively, if the gold price rises sufficiently, the miner may no longer be concerned about the foreign exchange risk.

Most barrier options are traded in the over-the-counter market where contracts are far from uniform. Disputes can sometimes arise over whether a *barrier event* has occurred, that is, the barrier has been hit or not. To avoid such disputes, transparency is needed regarding what constitutes a barrier event, for instance, whether it is based on quoted or transaction rates, the minimum size of the transaction required, the market hours during which it may occur and who can initiate the transaction. These issues are discussed in Hsu (1997).

To avoid a barrier event when the asset price only briefly touches the barrier, *Parisian options* require the asset price to be beyond the barrier for a certain length of time or a certain number of times before the barrier is triggered.

For a double-barrier option, it may matter in which order the barriers are hit, for example an upper knock-in barrier must be hit before the lower knock-out barrier is activated; that is, a down-and-out option knocks in only when the up barrier is hit.

Other types of exotics could knock in or knock out when the barrier is hit, for example a knock-in average rate option. Basket barrier options are also possible; both the barrier condition and the option component depend on the price of a basket of assets.

Clearly any of the partial and double features above can also apply to outside barriers, Parisian barriers or exotic barrier options. Each of the non-standard barriers may have a corresponding no-touch or one-touch version or an attached rebate.

The more complicated the payoff, the more likely a numerical method will be needed to price the option; a numerical method is the only practical way to price a window barrier or any more complex barrier option.

I.B.9.12.6 Relationships

Barrier options are cheaper than vanilla options

One of the main incentives for using barrier options for hedging is that they are *always* cheaper than the equivalent standard option. An extra constraint \mathbf{I}_A must be met before a payoff is made, making it harder to achieve a payoff. Since $\mathbf{I}_A \leq 1$, we always have

$$\mathbf{I}_A [\lambda(S(T) - K)]^+ \leq [\lambda(S(T) - K)]^+.$$

The greater the number of conditions to be met before a payoff can occur, the cheaper the option, for example:

- a double knock-out will be cheaper than either of the equivalent single up-and-out and down-and-out options;
- a partial knock-out option will be *more* expensive than the full-life equivalent, since it has less chance of knocking out;
- a partial knock-in option will be *less* expensive than the full-life equivalent, since it has less chance of knocking in.

In + Out = Standard

Either the barrier is hit or it is not hit before expiry. So at expiry the payoff of a knock-in option plus the payoff of the equivalent knock-out option is the same as the payoff of the equivalent standard option. The current price must also satisfy this relationship:

$$\begin{aligned} & \text{Price now of knock-in option} \\ & + \quad \text{Price now of equivalent knock-out option} \\ & = \quad \text{Price now of the equivalent standard option.} \end{aligned}$$

This also holds for double knock-out and double knock-in options: either one barrier is hit or neither is hit.

I.B.9.12.7 Ladders

In a one-rung ladder with strike K , if the asset price reaches a particular level H at any time during the life of the option, the exercise value at that level is locked in and the holder will receive at least this amount or the standard option payoff, whichever is greater:

$$V(T) = \max[\lambda(S(T) - K), \lambda(H - K)\mathbf{I}_A, 0]$$

where A is the event that S_t reaches the level H . The fact that a ladder depends on a level reached at some time during the life of the option suggests that ladders are related to barrier options and in fact a one-rung ladder can be constructed from barrier and standard options.

A ladder call with $H > K$ is equivalent to:

- (1) long a standard call, strike K ;
- (2) short a standard put, strike K ;
- (3) long a knock-out put, strike K , barrier H ;
- (4) long a standard put, strike H ;
- (5) short a knock-out put, strike H , barrier H .

If the asset price reaches H , the barrier options (3) and (5) knock out; the first two options give a synthetic forward at K . At expiry, if the asset price is greater than H the put (4) struck at H expires worthless, giving a payoff of

$$S(T) - K.$$

If the expiry asset price is less than H , the payoff will be

$$(S(T) - K) + (H - S(T)) = H - K.$$

If the asset price does not reach H , the knock-out options are still alive at expiry, so payoffs from the standard put (2) and barrier put (3) both struck at K will cancel, as do the standard put (4) and the barrier put (5) both struck at H , leaving only the payoff from the first option:

$$[S(T) - K]^+$$

Generalisations with more than one rung in the ladder can be constructed in the same way.

I.B.9.12.8 Lookback and Hindsight Options

Lookback options allow the holder to buy or sell the underlying asset at the best possible price achieved during the life of the option. The payoff of a lookback option is:

$$\text{Call:} \quad S(T) - \min_{\{0 \leq t \leq T\}} S(t),$$

$$\text{Put:} \quad \max_{\{0 \leq t \leq T\}} S(t) - S(T).$$

The asset is effectively bought at minimum asset price over the period for the call and sold at the maximum asset price for the put. For this reason they are sometimes called *no-regrets* options.

Hindsight options allow the holder to achieve the maximum return relative to a starting point. Their payoff is:

$$\text{Call:} \quad [\max_{\{0 \leq t \leq T\}} S(t) - K]^+,$$

$$\text{Put:} \quad [K - \min_{\{0 \leq t \leq T\}} S(t)]^+.$$

A natural choice for the strike is the asset price at the start of the contract: $K = S(0)$. In this case, the option will generally be exercised. They are also called *fixed-strike lookbacks*.

Lookback and hindsight options are understandably very expensive. It may also be difficult to monitor the absolute maximum or minimum reached by the asset price. To reduce the cost and make them easier to monitor, the continuous maximum or minimum can be replaced by a discrete version taken over a finite set of times T_1, \dots, T_n . The payoffs are the same as above with $\max\{S(T_1), \dots, S(T_n)\}$ replacing $\max_{\{0 \leq t \leq T\}} S(t)$ and $\min\{S(T_1), \dots, S(T_n)\}$ replacing $\min_{\{0 \leq t \leq T\}} S(t)$.

A hedging strategy involves buying a new at-the-money option each time a new extreme asset price is reached and selling the old hedge. However, such a strategy costs money at each rollover in the hedge, is impossible to maintain continuously and may generate large transaction costs.

I.B.9.13 Other Path-Dependent Options

I.B.9.13.1 Forward Start Options

A *forward start* or *stochastic strike* option has no strike price when the contract is initiated; the strike price is set at some later date T_1 , called the *strike-set date*, before expiry, that is $T_1 < T$.

The option is typically set to be at-the-money on the strike-set date but may also be set to be a proportion β in- or out-of-the-money. For example, $\beta = 1.05$ will set the strike price 5% above the asset price on the strike-set date; a put will be set 5% in-the-money and a call 5% out-of-the-money. The payoff of a forward start option is:

$$V(T) = [\lambda(S(T) - \beta S(T_1))]^+.$$

These options have very little delta until the strike is set; once the strike is set, the option becomes a standard option. So at the strike-set date, the delta could jump from close to zero to close to 0.50 when the strike is set at-the-money. Like compound options, they are sensitive to the future volatility over the period after the strike is set.

I.B.9.13.2 Reset Options

A reset option automates the decision to re hedge at a new level. In contrast to the forward start options of the previous section, a reset option does have a strike price K_0 at inception. At a series of subsequent dates T_1, \dots, T_n the option is reset to be at-the-money, if the asset price has moved to make the option out-of-the-money. So it will be more expensive than the equivalent standard option, and the more reset dates there are, the more expensive the reset option.

More precisely, at the i th reset date, the strike price is reset to:

$$K_i = \begin{cases} \min[K_{i-1}, S(T_i)] & \text{for a call} \\ \max[K_{i-1}, S(T_i)] & \text{for a put.} \end{cases}$$

The option payoff at expiry is that of a standard option based on the final strike price:

$$V(T) = [\lambda(S(T) - K_n)]^+.$$

We can rewrite the strike adjustment formula as follows:

$$K_i = \begin{cases} \min[K_0, S(T_1), \dots, S(T_i)] & \text{for a call} \\ \max[K_0, S(T_1), \dots, S(T_i)] & \text{for a put.} \end{cases}$$

This shows that reset options are closely related to discrete lookback options, discussed in Section I.B.9.12.8. In fact, a discrete hindsight option is a special case of a reset option in which the initial strike price is the asset price at the start of the option, that is, $K_0 = S(0)$.

As with many exotics, variations on this theme are possible, for example, a collar consisting of a bought call and a sold put, both of whose strikes reset as the asset price falls. The call is moving into-the-money at each reset, but the put is moving out-of-the-money. There is little business sense in such a put being sold alone, but it reduces the initial cost of buying the reset call.

The strike can also be reset by a proportion of how much the asset price has moved by the next reset date. The strike adjustment in this case would be:

$$K_i = \begin{cases} K_{i-1} - \beta[S(T_{i-1}) - S(T_i)]^+ & \text{for a call} \\ K_{i-1} + \beta[S(T_i) - S(T_{i-1})]^+ & \text{for a put} \end{cases}$$

The amount by which the strike may fall for a call or rise for a put may also be capped at some level B . This modifies the strike adjustment to:

$$K_i = \begin{cases} \max[\min(K_{i-1}, S(T_i)), B] & \text{for a call} \\ \min[\max(K_{i-1}, S(T_i)), B] & \text{for a put.} \end{cases}$$

I.B.9.13.3 Cliquet Options

It is not unusual in the world of exotic options for there to be a lack of consistency in how names of particular exotics are applied. For instance, in Zhang (1997) a *cliquet option* is defined as having a date T_1 , called the *cliquet date*, where the intrinsic value at that point can be ‘locked in’, even if the option subsequently moves out-of-the-money:

$$V(T) = \max[\lambda(S(T) - K), \lambda(S(T_1) - K), 0].$$

This kind of option occurs in executive options granted to senior management.

But the name cliquet option is also used (see Haug, 1997; or Ong, 1996) for a series of forward starting options in which the strike for the next exercise date is set to be to a positive constant times the asset price as at the previous exercise date:

$$V(T) = \sum_{i=1}^n [S(T_i) - \alpha S(T_{i-1})]^+.$$

So its price is the sum of the prices of the forward start options. This variety is sometimes called a *ratchet option*. In another variation (Lipton, 2003, Chapter 28), the payoff of a cliquet is given as:

$$V(T) = \sum_{i=1}^n \left[\frac{S(T_i)}{S(T_{i-1})} - K \right]^+.$$

These types of cliquets are very actively traded in the equity markets. Because of this liquidity and their sensitivity to volatility modelling assumptions, in Ayache *et al.* (2004) it is suggested that cliquets should be used as well as vanilla options for calibrating models that aim to explain the volatility smile (for example, jump-diffusion with stochastic volatility).

More complex structures referred to as cliquets are also popular in the equity markets, for example, a five-year *minimum coupon cliquet*⁸ whose payoff is a notional face value times

$$\max \left\{ \sum_{i=1}^5 \max \left[0, \min \left(\text{Cap}, \frac{S(T_i) - S(T_{i-1})}{S(T_{i-1})} \right) \right], \text{Floor} \right\}$$

with, for instance, Cap = 8%, Floor = 16%.

I.B.9.14 Resolution Methods

One reason for the success of the BSM model is that it is tractable, that is, calculations are relatively easy to perform. Within this model, many exotics have analytical pricing formulae. We have seen above that some seemingly ‘complex’ exotics, such as continuous barrier options, have analytical formulae while other seemingly ‘simple’ exotics, such as spread options, do not. Analytical formulae are very fast to execute, and speed in pricing is essential to perform risk management on portfolios of options within a practical length of time. If analytical valuation is not available for a particular exotic, analytical approximations are often developed. Formulae and approximations for common exotics can be found in Haug (1997) or Zhang (1997).

However, many exotic options, particularly with payoffs which depend on multiple asset prices or asset prices at many discrete points of time, can only be valued using a numerical method.

⁸ See Wilmott (2002).

Numerical methods frequently used to find option values include:

- binomial and trinomial trees;
- finite-difference solution of partial differential equations;
- Monte Carlo simulation;
- numerical integration.

It is beyond the scope of the PRM syllabus to describe these methods here. However, Chapter I.A.8 gives an introduction to binomial trees. Binomial and trinomial trees are ideally suited for pricing American options and other exotics in which the timing of an event affecting the payoff is at the holder's discretion. This is because we work backwards in time through the tree and so have knowledge of where the asset price will go. The binomial model is so widely used that an account of it is given in almost any introductory text on options, for example Hull (2003, Chapter 9).

If the payoff is path-dependent, for example it depends on the maximum or average, we need to keep track of which path through the tree has been taken. For an N -step binomial tree, there are $N+1$ possible final asset prices, but there are 2^N possible different paths through the tree. A method where we work forwards in time, like Monte Carlo simulation, is usually more efficient for such options.

Monte Carlo simulation can be used in general for any discretely path-dependent payoff, such as average or reset options. It can also be used for quite general modelling assumptions, including stochastic volatility or jumps in the asset price. Its major disadvantage is that it can be very slow to give an accurate valuation unless variance-reduction techniques are used. For example, it is more accurate to simulate the difference between an arithmetic average rate option payoff and the equivalent payoff based on a geometric average, then add the difference to the analytic valuation for the geometric average rate option. Variance reduction is a very effective way of making the simulation more accurate but can be technical to apply and specific to each option payoff.

Simulation is necessarily performed in discrete steps, so if a payoff depends on the asset price at every instant during the option's life, such as a continuously monitored barrier option, simulation can misprice the option unless techniques are used to minimise this discretisation error.

Black and Scholes (1973) originally derived their formula for valuing standard European options by showing that the option price satisfies a particular partial differential equation (PDE):

$$\frac{\partial V}{\partial t} + (r - f)S \frac{\partial V}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} = rV.$$

The finite-difference method is a technique for finding the solution V to the PDE by taking differences across the levels of a two-dimensional grid, or lattice, of asset prices and time steps. This can be a very efficient method for finding option values and is widely used in practice. Different exotic payoffs, such as barrier conditions, are accommodated by different boundary conditions at the edges of the lattice. Like binomial trees, the solution procedure proceeds backwards from expiry to the current time and so it is possible to value early exercise features. In more general models such as stochastic volatility models, a more involved PDE can generally be derived for the option value that can be solved using similar techniques.

Numerical methods can be applied very generally, to a wide variety of payoffs, using more general model assumptions than the BSM model. But no method works universally well for every exotic payoff. For an introductory account, refer to Hull (2003 Chapters 18 and 20); for more technical detail on their implementation, see Clewlow and Strickland (1998).

Pricing exotic options efficiently within models that are consistent with an implied volatility smile can be highly technical and is consequently beyond the scope of the PRM syllabus. A compilation of recent papers on state-of-the-art pricing and hedging techniques for exotic options can be found in Lipton (2003).

I.B.9.15 Summary

The purpose of exotic options is to match the hedging needs or speculative views of the end-user as closely as possible. Within the short space of this chapter, we have seen a wide variety of exotic payoffs. Starting from a simple cash-or-nothing payoff and strategies consisting of packages of vanilla and digital options, we have met options involving two assets, the most popular of these being spread options; options based on more than two assets, such as basket options; options in which one of the assets is priced in a foreign currency; options on other options; options where the holder can decide to exercise the option or change its strike price; path-dependent options, including average, barriers, lookbacks and cliquets. The different features in these exotics can be combined to arrive at even more complex products: a range note consists of a series of digital options; mountain options are path-dependent options on baskets of stocks.

Obtaining cost-effective hedging is a major motivation for the exotics user. This has made average rate options and barrier options, both cheaper than equivalent vanilla options, among the most popular exotics in the marketplace.

While exotic options, like more standard derivatives, transfer risk from the holder to the provider of the contract, the risks that remain for the market-maker are far from trivial. Despite the simplicity of the cash-or-nothing payoff, its discontinuity makes cash-or-nothing options exceedingly difficult to hedge. Likewise, reverse barrier options implicitly have a digital component that makes dynamic hedging virtually impossible. All options are exposed to changing volatility, particularly compound options, forward strike options and cliquets. Any option depending on more than one asset is exposed to assumptions regarding the correlation between the assets, an exposure which can seldom be hedged effectively.

While many exotics can be valued analytically in the BSM model, others can only be valued numerically. In this case they can also be very challenging to price within a reasonable period of time, particularly in models which take account of the random nature of volatility or potential jumps in the asset price.

The simple exotics dealt with in this chapter serve as an introduction to the creativity of the derivative markets in general. As new risks are identified, new products continue to be developed to hedge them. For instance, volatility and variance swaps have now been introduced to hedge the random nature of the volatility of financial assets; credit derivatives, designed to transfer credit risk from one counterparty to another, are also becoming widespread. It is clear that the inventiveness for new exotics in the financial markets is as yet far from exhausted.

References

- Ashraff, J, Tarczon, J, and Wu, W Q (1995) Safe crossing, *Risk Magazine*, July 1995.
- Ayache, E, Henrotte, P, Nassar, S, and Wang, X (2004) Can anyone solve the smile problem?, *Wilmott Magazine*, January, pp. 78–96.
- Black, F, and Scholes, M (1973) The pricing of option contracts and corporate liabilities, *Journal of Political Economy*, 81, pp. 637–654.
- Boyle, P P (1977) Options: a Monte Carlo approach, *Journal of Financial Economics*, 4, pp. 323–338.
- Clewlow, L, and Strickland, C (1998) *Implementing Derivatives Models*. Chichester: Wiley.
- Garman, M. (1989) Recollection in tranquillity, *Risk Magazine*, March, pp. 16–19.
- Harrison, J M, and Pliska, S R (1991) Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and Applications*, 11, pp. 215–260.
- Haug, E G (1997) *The Complete Guide to Option Pricing Formulas*. New York: McGraw-Hill.
- Hsu, H (1997) Surprised parties, *Risk Magazine*, April.

Hull, J C (2003) *Options, Futures, and Other Derivatives*, 5th edition. Upper Saddle River: Prentice Hall.

Jarrow, R (ed.) (1995) *Over the Rainbow*. London: Risk Books.

Lipton, A (ed.) (2003) *Exotic Options: The Cutting Edge Collection; Technical Papers Risk 1999–2003*. London: Risk Books.

Margrabe, W (1978) The value of an option to exchange one asset for another, *Journal of Finance*, 33(1), pp. 177–186.

Merton, R (1973) Theory of rational option pricing, *Bell Journal of Economics and Management Science*, 4, pp. 141–183.

Ong, M K (1996) Exotic options: the market and their taxonomy, in I Nelken, R Klein and J Lederman (ed.), *The Handbook of Exotic Options*. Chicago: Irwin/Probus.

Rubinstein, M (1991) Somewhere over the rainbow, *Risk Magazine*, November, pp. 63–66.

Rubinstein, M (1992) Exotic options. UC Berkeley, Walter A. Haas School of Business, Research Program in Finance, Finance Working Paper No. 220.

Wilmott, P (2002) Cliquet options and volatility models, *Wilmott Magazine*, November, pp. 78–83.

Zhang, P G (1997) *Exotic Options: A Guide to Second Generation Options*. Singapore: World Scientific.

I.C.1 The Structure of Financial Markets

Colin Lawrence and Alistair Milne¹

I.C.1.1 Introduction

This chapter provides a general overview of the global markets for financial securities and derivatives, examining the different ways they are organised and operated, and the arrangements that support them. Subsequent chapters (I.C.2–I.C.7) will examine in greater detail the markets for: money, bonds, foreign exchange, stocks, futures, commodities and energy.

Market structure is important to risk managers primarily through its impact on liquidity. A successful market, whether or not it exists in a physical location, brings together many buyers and sellers and is able to reduce search and transaction costs. Such a market provides the best possible conditions for financial risk management. Institutional arrangements for trading are crucial in achieving large volumes, good liquidity and thus promoting effective risk management.

Risk managers should concern themselves with market structure because it can also affect other aspects of risk, including credit risk, business risks, operational risk and basis risk. For example, trades on over-the-counter markets often have greater credit risk than those transacted on exchanges, but, because they can be tailored to individual requirements, they have less basis risk. An example of business risk is the capture of trading by electronic crossing networks, reducing the order volumes and revenues of NASDAQ dealer-brokers.

Section I.C.1.2 defines some important terms for market structure and provides an overview of the most important markets world-wide. Sections I.C.1.3 and I.C.1.4 discuss the main theme of this chapter, liquidity. They examine the key drivers of liquidity and the importance of liquidity for risk management, respectively. Section I.C.1.5 analyses the differences between trading on and off an exchange, while Section I.C.1.6 examines the impact of new technology on markets. The link between operational risk and market structure is made in Section I.C.1.7, which discusses post-trade processing. The intermediary role of brokers in markets is then the focus of Section I.C.1.8. Finally, new markets are discussed in Section I.C.1.9, while Section I.C.1.10 concludes.

¹ Colin Lawrence is Managing Partner of LA Risk & Financial Ltd, London, and Visiting Professor, Cass Business School, City University, London. Alistair Milne is Senior Lecturer, Faculty of Finance, Cass Business School, London. We are indebted to Carol Alexander and Elizabeth Sheedy for their comments but we remain responsible for all errors.

I.C.1.2 Global Markets and Their Terminology

Trading of financial securities, derivatives, and other financial contracts takes place in two settings: formal financial exchanges and more loosely organised over-the-counter (OTC) markets.

Financial exchanges are formalised trading institutions. Rights to trade are limited to members, and there are detailed and explicit rules governing the conduct of trade and the contracts or securities that are traded. Exchanges also collect and disseminate pricing information, and facilitate post-trade risk management and final trade settlement. OTC refers to any financial market transaction that does not take place on a formal exchange. The attraction of an OTC trade is that the buyer and seller are free to negotiate all the contractual details. But participants do not have the protection of exchange procedures and rules.

Most equity trading takes place through exchanges. An exception is NASDAQ, which is an OTC dealer market. The principle that liquidity is linked to market volumes can be illustrated in equity markets. Table I.C.1.1 presents statistics on the major equity markets. The bulk of equity trading, world-wide, is concentrated in a few major markets, mostly equity exchanges such as the NYSE, the London Stock Exchange, but also the OTC NASDAQ market. The six leading markets account for more than three-quarters of equity market capitalisation and equity trading.

We will now look at other major financial markets for debt, foreign exchange and derivatives, and examine how these activities are split between exchanges and OTC markets. Early markets for foreign exchange were the prototype for OTC arrangements. A market was established at a location where buyers and sellers of currency could approach established dealers and search for the best available exchange rate. Modern OTC markets rely on telephone and screen to link buyers and sellers with the market dealers, but the basics of the market remain the same. Dealers quote both bid and ask prices to prospective buyers and sellers. Dealers make a profit from order flow, both directly from the bid–ask spread, and also by anticipating short-term price movements using their privileged access to information on orders. But the dealer is also exposed to market risk from its holding of an inventory of currencies.

In such OTC markets a key role in the transfer of risk is often played by the ‘inter-dealer broker’, perhaps better referred to as the dealer’s broker– specialised firms where dealers may offload or purchase inventory. They, together with the continuous process of search by buyers and sellers, are the mechanism linking the market together.

Table I.C.1.1: Key statistics for the principal global equity markets

2002	Market cap (end year), \$trillion	Average daily turnover, \$billion	Average transaction value, \$thousand
NYSE	9.0	40.9	19
NASDAQ	2.0	28.8	12
London Stock Exchange	1.8	15.9	105
Euronext	1.5	7.8	31
Tokyo	2.1	6.4	n/a
Deutsche Börse	0.7	4.8	17
Other N. America	0.6	6.3	
Other European	2.1	10.1	
Emerging markets	3.0	13.1	
Total	22.8	134.0	

Source: World Federation of Exchanges

Dealers do still operate in some exchanges, notably the ‘specialists’ in the New York Stock Exchange, but, as we will discuss later, almost all security exchanges have found it more efficient to replace dealer market making with electronic order books.

Table I.C.1.2 shows statistics on the size of global debt markets. These are values of the issued principal, not market values, and therefore not strictly comparable with equity market capitalisations shown in Table I.C.1.1. Nevertheless they indicate that, taking account of government and financial institution debt (including the issue by banks of many kinds of asset-backed securities), the stock of debt on the market is considerably larger than that of equity.

**Table I.C.1.2: Stock of international and domestic market debt, end of September 2003
(US\$ trillion)**

	All maturities			Remaining maturity < 12 months		
	Domestic	International	Total	Domestic	International	Total
Government	18.1	1.1	19.2	4.1	0.1	4.2
Financial Institutions	15.5	7.8	23.2	4.0	1.4	5.4
Corporate Sector	4.9	1.4	6.3	0.6	0.2	0.8
Total	38.5	10.2	48.7	8.7	1.7	10.4

Source: Bank for International Settlements Quarterly Review

Table I.C.1.3 shows the market capitalisation of bonds listed on principal exchanges. This is only a fraction of total market debt, with a large concentration of issuance on the Luxembourg exchange, reflecting the importance of this listing for bonds traded in Europe. Much debt, notably all corporate and financial institution debt issued in the USA, is not exchange listed (all that is expected in this market is satisfaction of relevant SEC regulations).

Table I.C.1.3 Market capitalisation of bonds listed on principal exchanges (US\$ trillion)

	Luxembourg	London	Osaka	NYSE	Italy
Domestic Public Sector	0.0	0.0	3.6	1.1	1.1
Domestic Private Sector	0.1	0.1	0.1	0.2	0.1
Foreign	4.8	0.6	0.0	0.1	0.0
Total	4.9	0.7	3.7	1.4	1.2

Source: World Federation of Exchanges

With the exception of some government bonds, debt is traded on OTC markets rather than on exchanges. This is the case even when, as is usual in Europe, bonds are listed on an exchange. Issuers find it worthwhile to have a listing, demonstrating that they have satisfied certain exchange rules about accounting and other disclosure standards, even when trading itself still takes place outside of the exchange.

Table I.C.1.4 shows trading volume statistics on another major OTC market, that for foreign exchange. These figures are for daily turnover, during the month of April 2001, for spot transactions, forward transactions and foreign exchange swaps (the combination of spot and offsetting forward position treated as a single transaction). This is an example of a very liquid market; the value of total foreign exchange trading of these three types of contract is approximately 10 times the combined value of trading on the world's equity markets.

Table I.C.1.4: Daily foreign exchange turnover, April 2001 (US\$ billion)

Spot	387
Forward	131
Foreign Exchange Swaps	656
Total	1173

Source: BIS triennial survey

Table I.C.1.5 shows nominal and gross market values for OTC derivatives. Some, but not all, derivatives (including interest-rate swaps and the new market for credit derivatives) are OTC traded. Nominal contract values can be extremely large, notably for interest-rate swaps with notional outstanding principal of over US\$90 trillion. Gross market values (the sum of the absolute market value of contracts) is a better indication of the size and importance of these markets; the gross market value of interest-rate swaps, for example, is still small compared to the nearly \$50bn of outstanding market debt shown in Table I.C.1.2.

Table I.C.1.5: Major OTC derivatives, June 2003

	Notional amounts, \$US trillion	Gross market values, \$US billion
Interest-rate swaps	94.6	1126
Interest-rate options	16.9	434
Forward rate agreements	10.3	20
Currency forwards and fx swaps	12.3	476
Currency swaps	5.2	419
Currency options	4.6	101
Equity derivatives	2.8	260
Commodity derivatives	1.0	110
Other including credit derivatives	22.0	1083
Total	169.7	4029

Source: BIS quarterly review

Tables I.C.1.6a–6c show the number of contracts traded on the principal derivative exchanges. These tables illustrate the tendency for trading of particular contracts to concentrate on a small subset of exchanges. Thus equity index options are mostly traded on the Chicago Board Options Exchange, EUREX, and Euronext (Table I.C.1.6a). Government debt contracts are mostly traded on EUREX and the Chicago Board of Trade, while interest-rate contracts are mostly traded on EUREX and the Chicago Mercantile Exchange (Table I.C.1.6b).

Table I.C.1.6a: Principal exchanges for equity derivatives

Millions of contracts traded: 2002	Individual stock		Equity Index	
	Options	Futures	Options	Futures
AMEX	150.7		32.7	
BOVESPA	89.7		1.1	
Chicago Board Options Exchange	173.2		94.4	
Chicago Mercantile Exchange			5.4	212.2
International Securities Exchange	152.3			
EUREX	143.3	0.1	90.3	120.4
Euronext	323.6	7.6	108.3	51.9
Pacific SE	72.7		12.7	
Philadelphia SE/BOT	84.9		3.6	
World-wide	1300.0	57.1	420.2	531.3

Source: World Federation of Exchanges. Figures of less than 1 million contracts omitted from tabulation.

Table I.C.1.6b: Principal exchanges for fixed-income derivatives

Millions of contracts traded: 2002	Gov Debt		Interest rate	
	Options	Futures	Options	Futures
Chicago Board of Trade	54.7	205.6		7.7
Chicago Mercantile Exchange			105.6	202.5
EUREX	31.6	415.0		
Euronext		8.1	42.7	149.2
Korean Futures Exchange		12.8		
Sydney Futures Exchange	1.6	21.7		
Singapore Exchange				21.7
BM&F			2.5	48.6
World-wide	89.1	677.5	151.2	532.5

Source: World Federation of Exchanges. Figures of less than \$1m omitted from tabulation.

Table I.C.1.6c: Principal exchanges for currency/commodity derivatives

Millions of contracts traded: 2002	Currency		Commodity	
	Options	Futures	Options	Futures
Chicago Board of Trade			12.6	54.1
Chicago Mercantile Exchange	2.2	22.1	0.8	6.8
NY Mercantile Exchange			26.4	107.3
London Metal Exchange			2.3	56.3
Zhengzhou Commodity Exchange				14.6
Tel-Aviv SE	12.0			
BM&F	2.0	16.1		
World-wide	16.7	41.1	47.8	265.2

Source: World Federation of Exchanges. Figures of less than \$1m omitted from tabulation.

This feature, the concentration of trading in particular trading arenas, is also apparent when comparing OTC and exchange-traded derivatives. Prominent OTC derivatives such as interest-rate swaps are traded in a few centres (concentrated in London and New York), while there is no equivalent exchange-traded contract. The vast majority of currency derivatives are traded OTC (currency forwards dwarf currency futures).

I.C.1.3 Drivers of Liquidity

The previous section shows that trading activity tends to concentrate in particular markets. Why is this so?

Historically, financial markets have developed in particular locations, for example, in the coffee shops and alleyways of Venice, London and Amsterdam. Many of the earliest financial markets were for exchange of coin. Early markets for government bonds and shares operated in a similar fashion. The best prices and most reliable contracting, for both buyers and sellers, could be found in the established trading locations with greatest trading volumes. Liquidity – the promise of the best pricing with the least search effort – attracts more buyers and sellers. The presence of many buyers and sellers generates higher trading volumes which further narrows the gap between the prices for buying and selling (the ‘bid–ask spread’). This virtuous circle of liquidity/trading volumes has been a driving force for the development of financial markets, both historically and in the recent past.

As we discuss in this chapter there is a considerable variety in the way that markets are organised, from formalised exchanges to informal OTC markets (see Sections I.C.1.1 and I.C.1.5). There is also great variety in the mechanics of trading: in some markets computer technology matches buyers and sellers; while in others human beings still bring together market participants (Section I.C.1.6). But, although the technology of financial markets is often very different today from in the past, liquidity continues to be the primary driver of market development. Thus, while there is a general trend towards greater use of information technology to lower costs, a variety of different market arrangements can still succeed in capturing liquidity.

Automated trading systems are now common in both securities and derivatives trading (see Section I.C.5.5 and Chapter I.C.6). Outside of the USA the great majority of equity trading has now shifted from dealer-based market making onto electronic order books, such as the London Stock Exchange SETS system. We describe the operation of such systems in Section I.C.1.6. Although these automated matching systems operate very differently from dealer markets, they still result in the same kind of trade-off between pricing and liquidity as in dealer markets. The amount of order flow placed on the order book is the key factor in determining liquidity, just as normal market size predicts liquidity in a dealer market. The greater the order flow, the more likely it is that large trades can be transacted without adverse price impact. Also, for illiquid stocks with relatively little order flow, the gap between limit-sell and limit-buy prices is usually comparatively large.

Occasionally one marketplace can succeed in capturing liquidity from another. The best-known recent example is when the Swiss-German derivatives exchange EUREX, taking advantage of its lower-cost automated trading system, captured the trade in German government bond futures contracts from the London LIFFE derivatives exchange (see Section I.C.1.6 for more detail on this episode.). Subsequently, even when LIFFE adopted automated trading technologies as efficient as those of EUREX, it was unable to recapture the liquidity in this market. The shift in market liquidity was permanent.

Differences in liquidity can be especially pronounced when comparing individual securities. Table I.C.1.7 compares two measures of liquidity for two dealer-traded shares, normal market size (NMS) and the bid–ask spread. Normal market size is the maximum size of transaction at which a dealer is prepared to transact at the stated bid (buying) and ask (selling) prices.

Table I.C.1.7: An example of dealer-traded liquid and illiquid stocks

Stock	Market Cap \$ million	NMS \$ thousand	Bid	Ask	Spread %
A – liquid	10,000	200.0	9.43	9.45	0.21
B – illiquid	1	0.1	2.50	3.50	33.00

While these figures are not for actual stocks, they are representative of highly liquid and illiquid stocks traded in the major trading centres. They illustrate the marked difference in transaction costs (as represented by the final column, the bid–ask spread as a percentage of the mid-price) between the two groups. The liquid stocks generally have larger market capitalization and also have larger normal market size for transactions (although these are a relatively small proportion of total market capitalisation).

Liquidity risk is also linked with perceived credit risk on individual securities. Compare markets for otherwise similar securities with different degrees of credit risk, for example BBB corporate bonds and AA corporate bonds. Credit risk is difficult to assess and so securities with lower credit risk are seen as more homogeneous and can be bought and sold more easily and in larger quantities without a substantial price impact.

This relative lack of liquidity is reflected in ‘credit’ spreads over risk-free government bonds. Credit spreads on BBB bonds are much higher than those on AA bonds, a difference that cannot be fully explained by the higher historical rates of default on BBB bonds. This difference in liquidity was particularly evident during the ‘flight to quality’ in 1998 following the Russian default, when spreads on liquid exposures (AAA and AA bonds) narrowed while at the same time spreads on BBB bonds substantially widened. This divergence underscores the importance of liquidity risk to a risk manager and of understanding how the organisation of markets can affect liquidity risks.

There is also a close link between funding and liquidity, which has the tendency of pushing trading activity and price formation from cash onto derivatives markets. This is apparent, for example, in the much greater depth and liquidity of the interest-rate swap market compared to government bond markets. Taking a position in interest-rate swaps involves no exchange of principal, and hence no funding beyond any initial margin. Taking an equivalent position in the cash markets would require a large amount of capital funding. Hence market participants taking short-term positions almost always prefer to transact on the derivative rather than the cash market, and so trading and liquidity and best pricing all migrate onto the derivative interest-rate swap market (and other exchange-traded fixed-income derivative markets). A similar

phenomenon can be observed in the greater liquidity of many credit derivatives, relative to underlying corporate or sovereign securities. This liquidity advantage is the reason for transacting on theoretically redundant derivative markets. Cash and derivative markets have a symbiotic relationship, with prices moving closely together and any remaining discrepancy in pricing reflecting differences in liquidity and transactions costs.

I.C.1.3.1 Repo Markets

Liquidity can also be created through the removal of credit (counterparty) risk, most notably in money markets through the use of the ‘sale and repurchase agreement’ or repo. The repo is short-term contract in which one party agrees to sell a security (most often a high-quality bond such as a government bond or AAA corporate bond) to another party (the lender) and then repurchase subsequently at a higher price. Any coupon or dividend payments are still paid to the original owner, not the temporary purchaser of the security. (For a more detailed descriptions of repo transactions than we are able to give here, see Steiner (1997) and Choudhry (2002).)

To take a simple example, a creditor might agree with a bank to sell a quantity of government bonds for \$100 and to repurchase them three months later for \$101. This is effectively a loan of \$100 with a quarterly interest payment of \$1 (equivalent to an annual interest rate of just over 4%). The advantage of borrowing via a repo instead of using a conventional loan is that, provided the borrower has acceptable securities to pledge, they can borrow at close to a risk-free rate of interest (risk is almost entirely removed by requiring that the market value of the bonds used as collateral be substantially above the amount borrowed, thus the bonds might have a market value of \$110 for this loan of \$100, a ‘hair cut’ of 10%).

In developed money markets repos can be conducted over the whole range of maturities, from overnight to many months. Because they significantly reduce counterparty exposure, repos are the preferred form of transaction for many money market participants. They are used by central banks to conduct their short-term borrowing and lending so as to make monetary policy effective (market operations). The reduction of counterparty risk and consequent interest-rate reduction makes repo the preferred transaction for corporate borrowers; only if they are unable to pledge high-quality securities for repo borrowing will they take up other forms of borrowing such as issue of commercial paper, the drawdown of bank lines of credit, or (in the case of financial institutions) interbank borrowing. Furthermore, there are a variety of repo and reverse repo transactions which vary with time, duration and the specific collateral that is pledged (including stocks, bonds and commodities). Repo financing offers an efficient vehicle for fund managers to finance their securities, thus enabling leverage.

Repos are also used for the ‘shorting’ of securities (since every repo involves as its counterpart a security loan). The way this is done is through a ‘reverse repo’, that is, acting on the other side of a repo contract, accepting the loan of a particular security, and then immediately selling that security.

Thus, for example, to short \$1m of bonds issued by Daimler-Chrysler for a period of 1 month, from 1 April to 30 April, a market participant would first lend, say, \$975,000 to some counterparty in return for a pledge of the \$1m worth of Daimler-Chrysler bonds. The Daimler-Chrysler bonds would then immediately be sold for \$1m. In order to unwind the repo contract a month later, the bonds would have to be repurchased at the end-of-month market price. The overall deal will make a profit or a loss, depending upon how much Daimler-Chrysler bonds have fallen or risen in price over the month. If the bonds have fallen a lot in value then it will cost relatively little to repurchase at the end of the month and the shorting position will make an overall profit.

The development of repo contracts has added to the liquidity of money markets and, by making it easier to ‘short-sell’ securities, has also encouraged the closer integration of cash and derivatives markets. Indeed, repos play a crucial role in the pricing of all fixed-income derivatives. For example, in pricing a total return swap or credit default swap, traders or dealers will have to hedge these positions. Financing of these hedges invariably involves repos. For example, if a dealer sells a credit default swap, a hedge could involve the shorting of the corporate bond which would be financed through a reverse repo. Futures prices on fixed-income securities depend critically on repos.² The repo is hence a valuable tool for the risk manager and, through increasing liquidity, has also had a beneficial impact on financial markets as a whole.

I.C.1.4 Liquidity and Financial Risk Management

This section discusses the management of liquidity risks in different markets. Before doing this it is helpful to think more carefully about the nature of liquidity. There are many definitions – we all know what liquidity is, yet find it surprisingly hard to define it. Perhaps the best definition comes from Nobel prizewinner James Tobin. Liquidity is defined by the ability to sell or buy a commodity or service at ‘fair market’ value. If you are selling your house, then it might take months for you to sell at ‘fair market’ value if there is a lack of buyers. If you sell it instantly, you might have to sacrifice the price at which you sell. This adverse price impact results from the illiquidity of the housing market.

In technical jargon, a liquid market exists when the seller (buyer) faces a perfectly elastic demand curve. This means that an unlimited quantity can be sold (bought) at the market price. In financial markets, this clearly is not the case. In reviewing broker screens, one can note bid–offer spreads (or limit orders on an order-driven system) and one is forced to ask: ‘How much can I sell or buy at the quoted price?’. We can define the ‘implicit’ amount that an intermediary will transact as the normal market size. For example, the normal market size in \$/€ might well be around \$100m. What happens if a dealer has to sell, say, \$1bn? Quite clearly the screen’s bid–offer spread will no longer be relevant since other wholesale buyers will be unwilling to carry such a large inventory – they would have to line up buyers on the other side of the trade to whom they can sell on the dollars.

For a risk manager a major concern is the possibility of a sharp drop in the volume of order flow and hence in market liquidity. Unfortunately, such a dislocation in liquidity often results when unexpected information reaches the market and coincides with an increase in the volatility of market prices. For the risk manager this has important consequences for hedging. In addition, traditional value-at-risk estimates may understate the true risk of loss, especially in cases where large inventories (relative to normal market size) are held. (See Chapter III.A.2 for a discussion of value-at-risk models.)

One way of addressing this issue is to calculate the ‘liquidity adjusted’ value-at-risk, a measure that formally accounts for the impact of reduced liquidity (see Lawrence and Robinson, 1997). Liquidity-adjusted value-at-risk assumes that liquidity is ‘endogenous’. That is, liquidity is affected by the actions of the trader himself. We consider briefly how to model this aspect of liquidity risk.

In assessing liquidity-adjusted value-at-risk, this model examines the optimal speed with which to close a position. The inventory liquidator is faced with a trade-off: if he tries to unwind rapidly, he will be forced to pay a sizeable transaction cost due to the market impact and increased bid–ask spread associated with a large transaction. On the other hand, if he holds his position and liquidates slowly he will be exposed to adverse market movements for longer, leading to ongoing hedging costs and capital requirements to support the risk position. The model identifies the optimal pace of liquidation which occurs when the transaction costs (associated with that pace of liquidation) are just equal to the marginal cost of hedging plus the additional capital charge for holding on to the inventory. Simple value-at-risk models which implicitly assume that financial assets can be bought or sold at infinite elasticities can underestimate the value-at-risk dramatically, especially in more exotic markets.

² See for example, Choudhry, pp 331 -408 for an extensive analysis of the implied repo rate

This model of liquidity-adjusted value-at-risk suggests some practical conclusions about the relationship between liquidity and risk in different financial markets. Exposure to risk will depend upon the aggregate position held in a particular security or contract, and how this compares to typical trading levels. Liquidity risk can arise in even deep and liquid markets if the aggregate position is large enough so as to be difficult to unwind. Under stressed volatile conditions dealers may be less willing to deal and the ‘normal market size’ can itself fall, further exacerbating exposure to risk.

Finally, risk managers need to be aware that the ability to buy and sell at a fair market price can sometimes almost entirely disappear; because of the intimate links between cash and derivatives markets, this can have widespread impacts across financial markets. During the crises of both 1987 and 1998, it became difficult or even impossible to transact on many key markets. In 1987 both trading capacity and systems for posting margins in equity future markets were overwhelmed by the dramatic fall of equity prices, resulting in the effective closure of the derivative market and a consequent massive loss of liquidity on the cash equity market. In 1998 the massive positions taken by LTCM, and by other traders imitating their strategies, became totally illiquid as credit and liquidity spreads widened in the wake of the Russian default. The result was a collapse of liquidity in all but the most standardised products.

I.C.1.5 Exchanges versus OTC Markets

Why are some contracts traded on exchanges and others in less formal OTC markets? Financial exchanges offer their members a bundle of related services:³

1. Setting standards for traded financial products
2. Providing price information
3. Protecting against the risk of an agreement not being fulfilled (counterparty risk)
4. Facilitating the matching of buyers with sellers at agreed prices

We now discuss the first three of these services and how they are also supplied in OTC markets. We leave discussion of the matching of buyers with sellers until the following section on technological change.

In order for a company to list on a security exchange such as the New York Stock Exchange, the London Stock Exchange, or the Deutsche Börse, it must satisfy additional requirements over and above those of general company law. Accounts must be prepared according to specified standards and released at specified frequency (for larger companies quarterly statements are now

usually required). Companies are also required to publicly release any significant price information. All this gives greater confidence, to the purchaser of a listed equity, that the characteristics of the share are well understood and that there will be a ready market should there be a need to sell the share. Exchange rules also govern the market for corporate acquisition, imposing rules for the announcement of bids and the conduct of a contested acquisition.

The provision of price information is a major source of exchange revenues. A live feed of the current trading prices is a valued trading resource; hence, financial institutions and independent traders are prepared to pay substantial charges for live price feeds. Delayed feeds – of 15 minutes or so – are of little value to traders and can be obtained for free from websites and other sources. Trading prices are also needed, for example, by asset managers or hedge funds as a check that they are obtaining best execution from their brokers.

Price information on OTC markets emerges from the process of comparing quotes from several competing dealers. Provided the market is liquid, it is not difficult to obtain this information. Foreign exchange markets provide a good illustration, where there is such a high level of competition that quotes from different dealers differ by only very small amounts and a single market price emerges. The market for interest-rate swaps is similarly highly liquid. Indeed, so liquid is this market that the interest-rate yield curve emerging from swap transactions is regarded as a much more accurate measure of market interest rates than the relatively illiquid government bond curve. Government bonds are usually only traded actively fairly close to the time of issue. Hence the interest-rate swap market, where all maturities are traded actively all the time, has become the benchmark for interest-rate measurement.

Protection against counterparty risk is of particular importance in derivative contracts, which can be in force for several months or even years before they are finally settled. In contrast, the contract for the sale of a security is typically settled within 3 days. Derivative exchanges deal with this through the device of the *clearing house*, which becomes the counterparty to all derivative trades on the exchange, and imposes margin requirements on participants based on their net position *vis-à-vis* the clearing house. The question then arises, when will market participants choose to transact on an exchange and when will they choose an OTC market? In order to attract participation the major OTC markets have to have their own arrangements to deal with counterparty risk. They are open to highest-quality credits; only well-known financial institutions with credit ratings of AA or better are accepted. As a result, trade is possible without there being the same degree of netting or margining as is applied to exchange-traded derivatives.

³ See Lee (1998) for a detailed survey of the function and governance of exchanges.

This is well illustrated in forward foreign exchange where there are sufficient high credit quality participants that the OTC market has no difficulty providing the same control of counterparty risk and much greater liquidity than the competing exchange-traded contracts. As a result, the quantity of trade in OTC currency forwards dwarfs the liquidity of exchange-traded currency futures. The quality of entrants and size of transactions is different. There is no need for the same level of margining or other techniques to control counterparty risk, since the barrier to entry (the required credit standing) acts in place of ‘margining’.

OTC markets have developed their own procedures for obtaining contractual certainty and reducing counterparty risk. The International Swaps and Derivatives Association (ISDA) has developed a number of ‘master contracts’ covering the range of OTC derivatives. These master contracts allow for greater flexibility than anything traded on a derivatives exchange; buyers and sellers are free to alter specific aspects of the contract, to meet their own requirements. ISDA master agreements also support bilateral netting arrangements that act to reduce counterparty exposures.

Another key issue in the choice between exchange and OTC market is the trade-off between tailor-made solutions and basis risk. Exchanges provide liquidity in a few standardised contracts, but such contracts may not be appropriate for hedging purposes because small differences in maturity or other contractual details can lead to an unacceptable mismatch between the hedge and the position being hedged (basis risk). Much of the demand for interest-rate swaps arises from the fact that they can be tailored exactly to the hedging requirements of the participant, for example when replicating a structured interest-rate product. This explains why interest-rate swaps are an OTC rather than an exchange-traded product.

I.C.1.6 Technological Change

Changing technology is closing the gap between exchange and OTC trading. This section will review the impact of new technology on the process of matching buyers and sellers in both exchange and OTC trading. We then discuss how this is eroding the contrast between the two forms of market organisation (for an in-depth discussion, see Allen *et al.*, 2001).

Exchanges are well placed to take advantage of new technology to lower trading costs. They can invest, on behalf of members, in screen-based computerised systems for matching buyers and sellers. These so-called electronic order books replace the traditional floor-based, ‘open outcry’ trading mechanisms. An example is the SETS system used for share trading on the London Stock Exchange. Until the mid-1990s, share trades in London and elsewhere were executed on the floor

of the exchange; members of the exchange had to physically go onto the floor and find a matching buyer or seller. This matching was facilitated by identified locations for trade in specific shares. Dealers would congregate in the relevant location to perform their market-making function.

By moving to screen-based trading, SETS has achieved considerable reductions in the human resources involved in matching buyers and sellers, while also facilitating price comparisons. In order book systems of this kind, buyers and sellers enter an order onto the system for a given share, indicating both an amount and a price at which they are willing to buy or sell. For example, an order might be placed to buy 1000 Vodafone shares at £10.00 or less per share. If there is already an order in the system to sell 1000 Vodafone shares at £9.95 then an immediate match can be made for a trade of 1000 shares and the deal is executed at £9.95. If there is no matching order then the new order is placed in the system until one can be found.

Exchange members can view the SETS screen and choose to 'hit' available orders. They can see the whole range of orders, buy and sell, for a particular stock and the difference between the lowest offered sell price and the highest offered buy price. The major European exchanges, including Deutsche Börse and Euronext, operate using similar order book systems.

While the balance of advantage nowadays seems to lie with the electronic order-driven system for the matching of most trades there are important exceptions. For smaller less liquid equities there may be insufficient orders on the book. For this reason the London Stock Exchange also offers an alternative order-driven matching system for smaller, less liquid shares, known as SETS-m. This is a cross between a dealer quote-driven and automated order-driven system, with dealers providing regular buy and sell orders to maintain liquidity. The liquidity argument is also used to justify the continuation of open-outcry as the basis for trading on the New York stock exchange, the only major equity exchange that does not operate an electronic order book system; but a number of commentators believe that this resistance to computerised order matching technology reflects the vested interest of the exchange market makers, the 'specialists' who continue to profit from restrictive regulations that prevent orders being matched using lower cost mechanisms.

Another dramatic illustration of the impact of technology in reducing trading costs is in the field of exchange-traded derivatives. For over a hundred years, since the creation of commodity derivative exchanges in the late nineteenth century, the technique used to match buyers and sellers of derivatives has been 'open outcry', with buyers and sellers congregating in a restricted location, the trading pit, in order to buy and sell futures and options. Within the major exchanges there were separate pits for each category of derivative (e.g. a pit for Eurodollar interest-rate

options, a pit for dollar-sterling currency futures, etc.). Orders had to be taken into the pit by specialised runners who would then search amongst dealers in the pit to obtain the best price.

The advantage of low-cost computer technology is dramatically illustrated by the success of the Swiss-German-owned EUREX exchange in capturing trade in the major German government debt derivative contracts from the London-based LIFFE exchange. EUREX was able to offer much lower trading costs than LIFFE through the use of a computerised trading system instead of open outcry. While it took time to establish EUREX as a serious competitor, within a few months from mid-1997 to mid-1998, EUREX captured the liquidity in the major government bond future contracts (like the Bund contract) and this trading moved out of LIFFE onto EUREX. Subsequently open outcry trading was abandoned in LIFFE (now part of the Euronext group).

Open outcry trading still takes place in the major US derivative exchanges: the Chicago Board of Trade and the Chicago Mercantile Exchange, but is subject to challenge. In a period of only four years, the IESE electronic exchange has succeeded in capturing much of the liquidity in equity options (see Weber, 2003). EUREX has now established itself in Chicago as EUREX US, offering contracts that compete with some of the major contracts on the established Chicago exchanges. It is seeking, once again, to use a more efficient trading technology to capture market liquidity (see Young, 2004). The Chicago exchanges are responding, albeit rather slowly, with improvements in their own trading efficiency and a gradual switch to their own computerised systems.

Information technology is also having a major impact on OTC markets. Here the most notable example is the growth of electronic crossing networks (ECNs) such as Island which are capturing a large part of the volume of NASDAQ trading. OTC markets such as NASDAQ cannot coordinate technological change in the same way as organised exchanges. But participants will still look for better pricing (narrower bid–ask spreads) than is available from dealers.

ECNs are independent matching systems, working in a similar fashion to the electronic order books used on the European securities exchanges. The technology concerned is fairly cheap and now fairly standardised. The key feature of the ECNs is that they neither set prices nor carry inventory. Traders in NASDAQ shares can place orders on ECNs, at prices set with reference to quoted NASDAQ prices, and obtain a narrower bid–ask spread than is available from a NASDAQ dealer. As a result, much of the standard smaller size trade has migrated onto these electronic systems. Some 85% of NASDAQ orders are now said to be fulfilled electronically, with only the larger deals, which have price impact, being fulfilled by NASDAQ dealers.

As a result, there has been an effective convergence of the organisation of NASDAQ – the OTC securities market – with other exchange-based securities markets using electronic order books. In exchanges in Europe such as London and the Deutsche Börse, while orders of less than normal market size can be easily accommodated on the SETS system, larger orders can only be fulfilled by a process of negotiation. That is, the larger orders which have price impact are also routed away from automated electronic markets and directed to a phone- and screen-based dealer network. Only in the NYSE, where the electronic dealing technology is not yet available, is this approach not used.

This convergence may have some way to go. It can, for example, be argued that the abolition of exchanges and their replacement by OTC arrangements supported by ECNs could substantially reduce trading costs, in Europe and elsewhere (Domovitz and Steil, 2002).

Similar convergence can be observed between the markets for OTC and exchange-traded derivatives. Here new netting and margining services are being created for OTC derivatives, by, for example, London Clearing House-Clearnet with its SwapClear service. This service provides multilateral netting and margining for interest-rate swap contracts. Developments such as these allow OTC markets to provide as high a level of counterparty protection as exchanges.

It is sometimes argued that a switch from order matching by people (in pits or dealer markets) to electronic matching by machines can lead to a loss of liquidity and so increase both trading costs and risks. This kind of argument is made most often by floor-based traders defending themselves from the threat of technological change, but is voiced also by other more disinterested parties. As we have already seen, it is the case that purely electronic matching systems have difficulty in providing liquidity for larger trades. But it is a mistake to view such systems in isolation; in practice they are always accompanied by a range of alternative mechanisms for the execution of larger trades, including parallel dealer markets and wholesale brokerage. There is little evidence that electronic trading arrangements have difficulty in providing market liquidity on a day-to-day basis; indeed, the success of EUREX and IESE in capturing trading volumes and liquidity suggests the opposite. It is possible that in a severe market crisis, such as that of 1987, a market organised around an electronic trading system might suffer an (even) greater price decline than a pure dealer-based system. But this is a rather theoretical argument. In practice, commercial logic seems to be forcing all systems to adopt electronic systems. Management and regulators must simply focus on how to make these systems work as well as possible in stressed situations.

I.C.1.7 Post-trade Processing

After a trade is ‘executed’ further steps are required to complete the transaction. Nowadays, with increasing competition in trading spaces and consequent reduction in ‘bid–ask’ spreads, it is often such post-trade processing that creates the largest part of the cost of trading.

We can begin with the example of a securities trade, such as a bond or stock. After a deal is struck, three further steps are then undertaken:

Comparison and confirmation. Before the trade can be further processed it is necessary to conduct both comparison (does the information recorded by both sides of the trade agree?) and confirmation (is this what the investor really intended?). Following the trade, buyer and seller exchange messages confirming both their agreement to trade and all the details of the trade (security or contract, quantity, price, arrangements for settlement, etc.); and a broker needs to obtain positive confirmation from the investor (sometimes referred to as affirmation) that the trade complies with the original order.

Netting. A considerable reduction in the value and volume of securities trades for settlement can be achieved by netting offsetting cash and security flows. This is usually undertaken through a central counterparty (discussed further below). Netting allows a party’s commitments to be reduced to a single daily payment and single net figure for the acquisition of each security.

Settlement. The positioning of securities and the arrangement of payment occur prior to settlement. Settlement involves the final transfer of ownership of securities (delivery) in exchange for corresponding payments.

These steps take place sequentially. In most major financial centres most securities trades are processed on a ‘T+3’ basis, with matching and netting of trades completed by the end of T+1 (i.e., one day after the date of trade), preparation for settlement during T+2, and final settlement on T+3. Nowadays in all leading centres settlement is delivery versus payment (known as ‘DVP’), removing any risk of the loss of principal through counterparty default.

A number of institutions are involved in post-trade activities. Where netting takes place, it is usually on a multilateral basis and organised through a central counterparty or clearing house, such as London Clearing House-Clearnet (in UK, France, and a number of other European countries), EUREX Clearing (in Germany), or DTCC (in the USA). As we have already discussed, these central counterparties are extending their netting and risk management services into new markets.

Security settlement can be undertaken either by a central security depository (such as Euroclear-Crest, DTCC or Clearstream) or by a competing custodian bank. In order for a securities trade to settle it is necessary for the security to pass from one 'security account' to another and for there to be a corresponding payment from one bank account to another. Securities accounts can be offered either by the central depositories where securities are located or by custodian banks that hold securities on behalf of final investors. While DTCC offers a single post-trade processing solution for US transactions, post-trade processing elsewhere remains fragmented. As a result trading, especially cross-border trades, continues to be both costlier and riskier in Europe than in the USA.

Operational problems in trading are often associated with post-trade processing. Trading failures are much more common when manual processing is required, with the attendant risk of human error. Fortunately, such operational events tend to occur independently. Failures in post-trade processing may be higher than desired, but there is a good deal of diversification over time, average losses are predictable, and with improvements in procedures and reporting the level of losses can be reduced.

The goal nowadays for many back-office processes is to achieve 'straight-through' processing, that is, for all the details of post-trade processing to be input in standard form (referred to as 'standard settlement instructions') at the time the trade takes place; and then for all subsequent post-trade processing to take place without manual intervention. The great advantage of straight-through processing is the major reduction in operational risks and the high costs of manual processing.

The industry still remains some way from full straight-through processing for many trades, especially those that are cross-border or non-standard. As a result, there continue to be considerable costs and high levels of operational risk in post-trade processing. Operational risks are especially high for cross-border trading, a major issue for the development of European financial markets where post-trade arrangements continue to be highly fragmented, and also for investors in emerging markets.

Both central securities depositories and custodians also offer a number of other commercial services. They assist clients (securities brokers, asset managers) with their cash and collateral management, ensuring that they have sufficient securities and cash on account to settle all trades. They also lend securities (see Section 1.C.1.3 for discussion of the reverse repo), so providing liquidity to the post-trade process and to derivative markets. Services to securities holders include tax payment and reporting, corporate actions, and marking of portfolios to market.

I.C.1.8 Retail and Wholesale Brokerage

Not everyone can deal directly on financial markets. Exchange trading is limited to members – always established professional firms. OTC markets have their own limits on participation: only AA-rated firms are accepted as counterparties, only certain minimum sizes of deal are acceptable, and only regular participants who have established their identities with dealers and know the specific trading conventions for that market are able to trade freely.

Effectively this means that smaller deals, ‘retail trades’, cannot be handled directly on the market. Someone – the broker – has to collect together a number of retail orders and be responsible for passing them through onto the exchange or to an OTC dealer.

Financial firms with direct access to the financial markets can therefore make additional revenue by acting as brokers, that is, taking and executing orders on behalf of customers who cannot themselves deal directly on the market. (In fields such as insurance brokerage is also used to refer to the matching of buyers with sellers, but the mechanism is the same – in this case there are sellers who do not deal directly with small clients, and so a broker is needed to bring them together.) Brokerage has in the past been a highly profitable activity for investment banks and other financial institutions, using their privileged membership of exchanges or their established position in OTC markets to turn a substantial profit. Profit margins were exceptionally large for retail customers. They can still be high for customers who are naive enough to walk into a high-street bank to make a modest trade in a share or bond, and are charged a considerably greater fee than is imposed by the exchange where the deal takes place. But over recent years falling costs of information technology have allowed many newcomers to act as brokers, and margins have fallen dramatically.

What we have described so far is the low-risk activity of ‘retail brokerage’, handling orders that are too small, or come from insufficiently creditworthy customers, to be placed directly onto the market. Many of our comments about operational risks and operational costs in post-trade processing apply also to retail brokerage. In the past there have been high levels of operational costs, such as repudiated orders, but such costs have been fairly predictable and the application of information technology is greatly reducing such problems.

There is another important and distinct brokerage activity known as ‘wholesale’ broking, which means handling orders that are too large to be placed directly on the market and breaking them up and splitting them between buyers. Wholesale broking of this kind is carried out both by investment banks and by a number of small specialised wholesale brokerage firms. These firms

take little direct credit risk, but potentially huge operational risks due to the large value of deals that they handle on behalf of clients.

The line between brokerage and dealing is a fuzzy one. Brokers attract orders to both buy and sell. Once they have a sufficient order flow they can cut costs, both by matching trades on their own books without ever going to the market (so-called ‘internalisation’) and by meeting orders out of their own inventory. Regulators and audit firms therefore pay close attention, to ensure that customers are still offered at least the best market price (‘best execution’) even when the order is not fulfilled on the market.

Nowadays a large proportion of securities and derivatives trading is driven by hedge funds. The bigger hedge funds are often in the position of placing orders that are too large to be placed directly onto the market. Therefore, for these funds a critical relationship is with their primary broker who will handle all aspects of trade execution, passing on orders and dividing them amongst different brokers for execution, as well as providing the entire range of post-trade reporting and analysis services. The primary broker is also a major secured lender to the hedge fund (making use of the repo contract described earlier). The primary broker to a hedge fund is always one of the major investment banks.

Primary brokerage has grown to be one of the more lucrative investment banking activities. It offers large margins and also the opportunity to observe hedge fund activities and make profitable proprietary trades based on knowledge of the client hedge funds’ positions.

I.C.1.9 New Financial Markets

This chapter has compared the activities on the principal securities and derivative markets. An analysis of the structure of financial markets would, however, be incomplete, without a discussion of the rapid development of new markets, especially for structured products, of the past decade.

Much of the trading of interest-rate swaps is related to their use for the creation of structured notes and in structured finance. Structured notes are OTC interest products customised to client specifications. Structured notes are extraordinarily flexible; they can be created with virtually any conceivable interest-rate profile, based on both floating and fixed rates, domestic and foreign currency. One well-known example is the inverse floating-rate note, where the return varies in the opposite direction to LIBOR interest rates. If LIBOR increases by 100 basis points, the interest paid on this product declines by 100 basis points.

Structured notes provide great flexibility in meeting client needs. They can be tailored precisely to match anticipated interest and currency exposures, that is to say, they are very effective hedging tools. They also can be used to get around regulatory restrictions (institutions barred, for example, from investing in emerging markets can still obtain the benefits of exposure to the higher returns available in emerging market products by purchasing a structured note from an OECD financial institution with returns linked to one or a basket of emerging markets). They are also very effective for tax purposes; for example, a structured note may be issued by a vehicle situated in a country with a double-taxation treaty with the country of the purchaser, allowing the reclaim of withholding tax, even though the financial returns are related to the interest rates in a country with no double-taxation treaty.

The more recent development in the field of structured finance has been the explosive growth of collateralised debt obligations (CDOs), the major structured credit product. Table I.C.1.8 shows figures on the volume of CDO issuance.

Table I.C.1.8: Rated CDO volumes (\$bn)

1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
0.2	1.3	1.4	2.5	18.7	38.1	81.6	120	160	200	250

Source: Bank of America and Moody's Investor Services, cited in Tavakoli (2003, p. 11).

CDOs are one of the modern vehicles through which loans, bonds or other financial assets can be securitised (see Chapter I.B.6). The innovation in these instruments is that they enable financial intermediaries to transfer or share the risks of these obligations with other investors. The CDOs are a tradable instrument with a range of debt issuance through cascading tranches of risk.

Important drivers of this new market include the introduction of the euro, which supported a sufficiently large and liquid market in euro-denominated credits, which could not have been supported in, say, Deutschmarks or French francs alone. The development of the synthetic CDO has also been an important factor. With synthetic CDOs, bank exposures are not sold to the securitisation vehicle, but instead the credit risk is transferred through the use of credit default swaps or other credit derivatives. These credit derivatives can then be used as the backing for tradable CDO paper. This arrangement is very attractive to many banks because they can maintain relationships with their borrowers (loans are not sold), and because the CDO is delinked from the bank funding, it becomes a pure credit transfer. Some two-thirds of the 2002 issue of CDOs was in the form of such synthetics.

Another growth area has been the secondary loan trading market, which again has grown significantly. It enables those banks unable to participate in syndications to buy loan assets, allowing banks that are over-concentrated to reduce their exposure (see the website of the loan syndication and trading association www.lsta.org for statistics on the growth of this market).

The major impact of securitisation, together with secondary loan trading, is quite revolutionary. Traditionally the business model of the commercial bank was to take deposits and to lend and hold both these assets and liabilities on balance sheet. But today we are in a very different environment, with the huge growth of securitisations (including CDOs) enabling banks to package their loans and sell them as securities. Retail lending has for some time been transferred off-balance sheet, with the biggest market for asset-backed securities that for US mortgage-backed securities (backed by Ginnie Mae and Freddie Mac), which is bigger and often more liquid than the US government treasury market. Now the corporate loans are also no longer a long-dated 'buy and hold' instrument; as the CDO and secondary loan markets expand, corporate lending is becoming an increasingly liquid asset.

A third closely related development has been the growth of the credit derivative market, including credit default swaps enabling banks to 'insure' against defaults. A critical feature of the credit default swap, unlike a credit insurance contract, is that there is no requirement that the purchaser of protection hold the underlying insured assets. In the event of default 'event' the default swaps can be settled either by transfer of the physical asset or by payment of a cash difference, based on a post-default market price of the insured asset. The rapid growth of the credit derivatives market has however concerned regulators, especially the concentration risk of protection; with a few major institutions selling most of the protection.

These new markets create both opportunities and risks. We can expect the development of these new liquid markets for credit to improve the ability of risk managers to control and respond to credit risk and to reduce the overall costs of intermediation between savers and borrowers. But at the same time, because these markets are very new, there must be concerns about how well participants understand the risks they are taking on. The possibility of mispricing and of dramatic loss of liquidity in the event of crisis cannot be dismissed.

I.C.1.10 Conclusion

This chapter has surveyed the structure of financial markets. We have compared activity in securities markets (equities and bonds), foreign exchange and derivatives and discussed key organisational and structural features of modern markets.

We emphasise the association between the organisation of markets and the availability of liquidity, that is, the opportunity to buy or sell without affecting prices. We noted the ‘virtuous circle’ of liquidity, the tendency for market activity to concentrate on a single trading platform so as to obtain the best pricing. We also observed the increasing liquidity of derivative markets, the ‘symbiotic’ relationship between cash and derivatives pricing, and the tendency for trading activity to migrate from cash to derivatives.

One lesson is that effective management of liquidity risk requires an appreciation of how markets are organised and how they interact. Where contracts are homogenous, markets are effectively linked together, and positions are small, liquidity problems are relatively unlikely to arise; but where products are diverse, markets are fragmented, and positions are large then liquidity is a serious potential problem.

We have compared over-the-counter markets and exchanges, noting that the development of new trading technologies and the spread of central counterparty services is leading to a convergence of these two forms of market organisation. Standard trades are now often matched through automated systems – whether on exchanges or OTC – and in both cases dealers or brokers are often taking responsibility for matching larger orders.

While there are a few markets – notably the NYSE – where matching of buyers and sellers is still undertaken through dealer quotation, automated trade matching is now standard practice for standard size trades. Automation is significantly reducing trading costs, but liquidity for large trades still depends upon the intervention of brokers to match buyers and sellers. A brokerage service is also still used for small retail trades, but here competition and new technology have dramatically reduced both costs and profit margins.

We have briefly described post-trade securities processing, including both the netting of credit risks through central counterparties and the final transfer of ownership against payment. We have noted the potential for high levels of operational costs and risk wherever there is manual intervention. The use of information technology to standardise procedures (‘straight-through processing’) is reducing these operational costs but is still limited for many cross-border transactions.

The final section concluded with a review of some of the newest financial products, including structured securities and credit products such as loan trading and credit derivatives. These products, like previous financial innovations, promise to improve liquidity and reduce the overall costs of financial intermediation between savers and investors. But because they are so new these products are not always well understood, and the possibility of losses from mispricing or a systemic crisis cannot be ignored.

References

Allen, H, Hawkins, J, and Sato, S (2001) Electronic trading and its implications for financial systems. In S Sato and J Hawkins (eds), *BIS Papers No. 7 – Electronic Finance: A New Perspective and Challenges* (available via <http://www.bis.org>).

Choudhry M, (2002), *The Repo Handbook*, Butterworth Heinemann,

Domowitz, I, and Steil, B (2002) Innovation in equity trading systems: the impact on trading costs and the cost of equity capital. In B Steil, D G Victor and R R Nelson (eds), *Technological Innovation and Economic Performance*. Princeton, NJ: Princeton University Press.

Lawrence, C, and Robinson, G (1997) 'Liquidity, dynamic hedging derivatives and value at risk. In *Risk Management for Financial Institutions: Advances in Measurement and Control*. London: Risk Books, pp. 63–72.

Lee, R (1998) *What is an Exchange?* Oxford: Oxford University Press.

Steiner, R (1997) *Mastering Repo Markets*. London: FT Pitman.

Tavakoli, J M (2003) *Collateralized Debt Obligations and Structured Finance*. New York: Wiley

Weber, Bruce (2003) Adoption of electronic options trading at the IESE. *IT Pro*, July (available via http://www.london.edu/sim/Working_Papers/SIM22.pdf).

Young, P L (2004) The Battle of Chicago: Chicago after Eurex US. In *The Handbook of World Stock, Derivative and Commodity Exchanges 2004* (available via <http://www.exchange-handbook.co.uk>).

I.C.2 The Money Markets

Canadian Securities Institute¹

I.C.2.1 Introduction

The interest-rate market is the market in which individuals and businesses lend cash to other individuals and businesses and in return receive compensation in the form of interest. Most lending and borrowing occurs through some sort of intermediary, including deposit-taking institutions such as banks, trust companies or credit unions, and investment banks and dealers who underwrite interest-rate securities on behalf of borrowers, including corporations, governments, and supranational institutions such as the World Bank.

The interest-rate market is not a market with a single physical location. It is made up of several thousand financial institutions and several million customers who interact face-to-face, by telephone, over the Internet, or on private computer networks every business day around the globe. Lenders constantly seek out the highest interest rates, while borrowers continuously search for the cheapest source of funds.

When a lender makes a loan to a borrower, an interest-rate asset is created for the lender and an interest-rate liability is created for the borrower. This chapter examines the various instruments that are available in the interest-rate market, broken down into the market for deposits and loans and the market for securities.

I.C.2.2 Characteristics of Money Market Instruments

The cash market for interest-rate assets and liabilities can be thought of as two distinct but related markets: the market for *deposits and loans* and the market for *securities*. The wide variety of instruments in these two categories, which are known collectively as *fixed-income instruments*, may have any or all of the following six characteristics:

- *Term*. The term of a fixed-income instrument is the length of time that the borrower borrows the money. The day that marks the end of the term is sometimes referred to as the maturity date. Terms can range from as short as one day to an infinite amount of

¹ Canadian Securities Institute, Toronto, Canada.

time.² Most terms, however, are between one day and 30 years. Some instruments, known as revolving loans, do not have a specific maturity date; the borrower may pay off a revolving loan at any time.

- *Principal.* The principal is the amount that the borrower agrees to repay to the lender on the maturity date. The principal has many synonyms in the interest-rate market, including face value, par value, and deposit or loan – although technically the last two terms describe a type of instrument, rather than the amount lent. Some instruments require the borrower to repay the entire principal on the maturity date, while others require the borrower to repay the principal in instalments over the life of the loan. The principal is usually the same as the amount lent, although in some instances it may be more.
- *Interest rate.* This is the amount that the borrower agrees to pay the lender for the use of the money, usually expressed as an annual percentage of the principal. Interest payments may be paid in regular instalments over the term of the instrument, or in a lump sum at the maturity date. For example, an interest rate of 6% per annum means either that the borrower must pay the lender interest of 6% on the value of the principal every year, or that the borrower must pay a proportion of 6% if interest payments are made more frequently, or that the borrower must pay the equivalent of 6% in annual interest at maturity when the principal is repaid. For interest-rate securities such as bonds that pay interest at regular intervals, the interest rate is also known as the coupon rate. The interest rate can be fixed over the term of the loan, or can vary, or float, according to market interest rates. When the interest rate varies, there may be a limit placed on how high or low the interest rate can go.
- *Marketability.* If ownership of the interest-rate asset or liability can be readily transferred to a third party, the instrument is said to be marketable. Most interest-rate securities are marketable instruments.
- *Security.* Fixed-income instruments can be either secured, or collateralised, by specific assets, or unsecured. If an instrument is secured, the lender has the right to take ownership of the specified assets if the borrower does not fulfil his or her obligation to repay the principal. If an instrument is unsecured, the lender has no specific claim to any of the borrower's assets if the borrower does not repay the principal. In this case, the

² In the 1980s some companies issued debt known as perpetuities, which had no stated maturity date. The perpetuities called for the issuing companies to pay interest for ever. Almost all of the perpetuities, however, included a provision to allow the issuer to redeem them after a certain number of years.

lender is an unsecured creditor who relies on the borrower's general ability to meet financial obligations.

- *Call or put features.* Some instruments have a provision that allows the borrower to repay the principal (a call feature) or the lender to demand repayment of the principal (a put feature) before the maturity date. See Chapter I.B.8 for more details.

I.C.2.3 Deposits and Loans

Banks, credit unions and trust companies (which as a group will be referred to as just 'banks') and large investment dealers are the dominant players in the domestic markets for deposits and loans to individuals and businesses.

I.C.2.3.1 Deposits from Businesses

Banks accept deposits from businesses in three basic forms.

1. *Demand deposits*, more commonly known as chequing accounts, can be withdrawn by the depositor at any time, without giving any notice to the bank. Most banks pay little or no interest on demand deposits.
2. *Notice deposits*, which consist primarily of savings accounts, require the depositor to give the bank advance notice before withdrawing the funds, although this requirement is rarely, if ever, enforced. Notice deposits are floating-rate deposits. The rates offered by most banks are usually quite low, and they change infrequently.
3. *Fixed-term deposits* have fixed terms and must be repaid, with interest, to the depositor on the maturity date. Fixed-term deposits are also known as term deposits or time deposits. Some fixed-term deposits, however, have a provision that allows the depositor to withdraw the deposit before the maturity date. There may or may not be a penalty for doing so.

Banks offer fixed-term deposits with both fixed and floating rates of interest. The interest rate on floating-rate term deposits is usually tied to a benchmark interest rate, also called a reference rate. These rates include administered rates, which are set by various institutions or policy makers, such as a bank's prime rate, or a market rate, which is determined by the trading activity in a certain market, such as the 90-day Treasury bill rate.

I.C.2.3.2 Loans to Businesses

Banks lend money to businesses in many different forms. Loans to medium-sized and large corporations are usually structured as *credit facilities*, also called *credit lines*. A credit facility is a flexible, customised arrangement in which the corporation has a variety of ways to borrow from the bank. Banking is a competitive industry, and individual banks try to win new business and retain their current clients by offering a wide menu of choices. Typically, credit facilities allow corporations to borrow through fixed-rate term loans, floating-rate loans, or bankers' acceptances, which are discussed in detail in Section I.C.2.4. Except for the most creditworthy corporations, most bank lending is secured by the corporation's physical assets or receivables.

Credit facilities are tailored to suit the needs of the borrower. At any given time, a borrower does not usually require all the funds that the bank makes available in different forms as part of the facility, but has access to the money whenever it is needed. To compensate the banks for being ready and willing to lend the entire amount of the credit facility, the corporation usually pays a standby or commitment fee on the facility's unused portion. The fee is expressed as a percentage of the unused portion, and is usually quite small.

Banks also make one-off floating-rate loans to corporations based on the bank's prime rate plus a spread. So-called prime-based lending is more expensive for corporations than, for example, bankers' acceptance issues, because a bank's prime rate is always higher than its bankers' acceptance rate.

The largest banks also participate in syndicated lending to very large, mostly publicly traded financial and non-financial corporations. In syndicated lending, a group of banks lend money on common terms to a single borrower. The loans can be structured as credit facilities, with a range of options for the borrower, or as one-time loans with fixed- or floating-rate terms. The total amount lent is usually quite large, and many of these loans exceed \$1 billion. By sharing the loan among several lenders, banks lessen their exposure to a given borrower.

I.C.2.3.3 Repurchase Agreements

A repurchase agreement, or *repo*, is a loan in which a borrower sells a security to a lender at one price with an agreement to buy the security back on a future date at a higher price. Although it is simply the other side of the same agreement, the lender's position in the agreement is generally referred to as a reverse repurchase agreement, or *reverse repo*.

An *overnight repo* is a repo with a term of one day. Term repos are repos with any term longer than one day. The higher price at which the security will be repurchased is determined by the repo

rate. The security sold to the lender acts as collateral for the loan. If the borrower cannot repay the funds on the maturity date of the repo, the lender gets to keep the securities.

The repo rate for a particular repurchase agreement depends on several factors, including the quality of the collateral and the term of the agreement. The quality of the collateral affects the credit risk and liquidity of the security. The effect of the repo's term on the repo rate depends on the general level of interest rates in the market for different terms to maturity.

Investment dealers make extensive use of repurchase agreements to finance their inventories of equity and interest-rate securities. They pay interest according to the standard formula for computing the interest on a short-term loan:

$$\text{Interest Paid} = \text{Principal} \times \text{Interest Rate} \times (\text{DTM}/\text{ADC}) . \quad (\text{I.C.2.1})$$

where DTM = the number of days until maturity,

ADC = the denominator of the relevant day-count convention.³

Example I.C.2.1

Suppose ABC Securities has just bought \$5 million worth of a Government of Canada bond from National Securities. The bonds will be held in inventory for resale to ABC's clients. ABC expects to sell the bonds within a week. To pay for the bonds, ABC can use its own funds or it can borrow the money in the repo market.

To finance the purchase, ABC enters into a seven-day term repo with CIBC. ABC agrees to deliver to CIBC the \$5 million worth of bonds in return for \$4,995,205.48 today. ABC will then use the proceeds from this loan to pay National Securities for the bonds.

At the end of the repo agreement, CIBC will return the bonds to ABC and ABC will give \$5 million to CIBC. The difference between the two payments, \$4,794.52, represents the interest on the loan. This was calculated based on a repo rate of 5% . Then, using (I.C.2.1) the interest paid is

$$\$5 \text{ million} \times 0.05 \times \frac{7}{365} = \$4,794.52.$$

³ A day-count convention is a method of calculating interest for periods of less than one year. The different segments of the interest-rate market use different day-count conventions, including actual/actual (or actual/365), actual/360, and 30/360, which assumes that each month is 30 days long and therefore a year has only 360 days. See Chapter I.C.3 for further details.

If ABC succeeds in selling the bonds to its clients by the end of the week, it will have the funds it needs to repay CIBC and will get the bonds back so that they can be delivered to the clients who purchased them.

I.C.2.3.4 International Markets

The international market for deposits and loans is known as the *Eurocurrency market*. This is the market for term deposits and loans in a currency other than the local currency of the bank branch that is accepting the deposit or extending the loan.

The Eurocurrency market, which is most active in London, England, grew rapidly in the 1980s, largely in response to regulatory restrictions placed on financial institutions within their domestic markets. Although some of these regulations have since been relaxed, many still exist, making the Eurocurrency market a vital source of relatively low-cost, short-term funding for large multinational corporations and internationally active financial institutions. The largest segment of the Eurocurrency market is by far the *Eurodollar market*. This is the market for US dollar time deposits and loans. Let us examine how a Eurodollar deposit is created.

Example I.C.2.2

Suppose that ABC Corp., a Canadian company, owns a maturing time deposit of US\$10 million with Citibank, which is based in New York City. ABC will not require the funds for another three months, so the company's treasurer solicits two quotes on a new three-month time deposit: one from Citibank in New York and one from the London branch of Deutsche Bank. Deutsche Bank offers ABC a slightly higher interest rate than Citibank, so the treasurer decides to deposit the money with Deutsche Bank in London.

When the original term deposit matures, ABC owns a US\$10 million demand deposit (i.e., a chequing account) with Citibank in New York. The treasurer of ABC instructs Citibank to transfer the demand deposit to Deutsche Bank's account with Citibank in New York⁴. Once this is complete, the Eurodollar deposit has been created. Table I.C.2.1 shows how the scenario looks using T-accounts⁵.

⁴ This assumes, of course, the Deutsche Bank has an account with Citibank in New York.

⁵ A T-account is an accounting tool used to keep track of two-sided transactions. In our example, assets (deposits in other banks or loans to other parties) are on the left-hand side and liabilities (deposits due to other parties) are on the right-hand side

⁶ Again, we assume that Royal Bank's London branch has an account with Citibank in New York.

Table I.C.2.1: Creation of a Eurodollar time deposit

UNITED STATES		LONDON	
Citibank		Deutsche Bank	
	\$10 million demand deposit due to Deutsche Bank	\$10 million demand deposit in Citibank	\$10 million Eurodollar time deposit due to ABC Corp.

Note that Citibank is simply transferring ownership of the demand deposit to Deutsche Bank. The demand deposit, and the actual funds backing it, remain in New York. In other words, there is no US\$10 million sitting in Deutsche Bank’s vaults in London.

But the scenario does not end here. Deutsche Bank now owns a US\$10 million demand deposit and has promised to pay interest to ABC Corp. on its three-month Eurodollar time deposit. Deutsche Bank needs to put that demand deposit to work in a manner that will generate enough interest not only to cover its interest payment to ABC, but also to turn a small profit. If it cannot quickly find a customer who needs to borrow US\$10 million, Deutsche Bank will lend the money to another bank in the Eurodollar interbank market. This process may continue for several more stages, with several banks borrowing the money and then relending it to other banks, until the demand deposit in New York is finally transferred to a borrower who actually needs the US\$10 million.

Suppose that Deutsche Bank buys a three-month Eurodollar time deposit from Royal Bank in London. Deutsche Bank will instruct Citibank in New York to transfer ownership of the demand deposit, this time to Royal Bank.⁶ To entice Deutsche Bank to make the Eurodollar deposit with it, Royal Bank will offer Deutsche Bank a slightly higher rate than Deutsche Bank has promised to pay ABC Corp.

Now suppose that Royal Bank does have a customer, XYZ Inc., that needs a US\$10 million loan, which Royal Bank is able to provide thanks to its newly acquired ownership of a US\$10 million demand deposit with Citibank. To complete the cycle, Royal Bank instructs Citibank to transfer ownership of the US\$10 million demand deposit to XYZ Inc.⁷ Table I.C.2.2 presents the new scenario.

⁷ Once again, we assume that XYZ has an account with Citibank in New York. If none of these Citibank-account assumptions were true, the demand deposit would be transferred between each party’s New York-based bank.

⁸ From ‘The BBA LIBOR Fixing – Definition’ on the British Bankers’ Association website, www.bba.org.uk.

In the above example we used a three-month deposit to illustrate how Eurocurrencies are created, but banks in the Eurocurrency market regularly accept deposits (and extend loans) with a range of terms, from one day up to one year. The most popular terms are one day, one week, one month, three months, and six months.

The type of interest that the banks pay on Eurocurrency deposits is known as *add-on interest*. That is, interest is added to the deposit amount and paid when the deposit matures. The actual amount of interest is calculated on a money market yield basis using (I.C.2.1) with an actual/360 day-count convention.

Table I.C.2.2: A second Eurodollar time deposit and the final borrowing

UNITED STATES		LONDON	
Citibank		Deutsche Bank	
	\$10 million demand deposit due to XYZ Inc.	\$10 million Eurodollar time deposit in Royal Bank	\$10 million Eurodollar time deposit due to ABC Corp.
		Royal Bank	
		\$10 million loan to XYZ Corp.	\$10 million Eurodollar time deposit due to Deutsche Bank
		XYZ Inc.	
		\$10 million demand deposit in Citibank	\$10 million due to Royal Bank

Example I.C.2.3

In the Eurodollar market, if a corporation puts US\$1 million into a three-month time deposit on 1 March with an interest rate of 6% per annum, the number of days for which the borrower will pay interest at maturity is 92, calculated as follows.

2 March to 31 March	30	days
1 April to 30 April	30	days
1 May to 31 May	31	days
1 June	1	day
	<hr/>	
	92	days

Note that the day count begins the day after the deposit is made and ends with the maturity date.

Using (I.C.2.1), the total interest received by the depositor at maturity for the 92 days is:

$$\$1 \text{ million} \times 0.06 \times \frac{92}{360} = \$15,333.$$

At maturity, the deposit will be worth the US\$1 million principal plus US\$15,333 interest.

I.C.2.3.5 The London Interbank Offered Rate (LIBOR)

On each business day in London, banks in the Eurocurrency interbank market constantly lend and borrow Eurocurrency deposits to and from each other. The most popular currencies represented include the US dollar, the Japanese yen, the Swiss franc, the Canadian dollar, the Australian dollar, and the euro. The rates the banks bid for deposits and offer for loans in each of these currencies are regularly posted on quotation and automated trading systems such as Reuters, Bridge Telerate, and Bloomberg.

At 11:00 a.m. London time each day, the British Bankers' Association (BBA) surveys the rates offered by at least eight banks chosen for their 'reputation, scale of activity in the London market, and perceived expertise in the currency concerned, and giving due consideration to credit standing.'⁸ It ranks the quotes from highest to lowest, drops the highest and lowest 25%, and takes the average of the remaining 50%.⁹ The result is the official BBA *London Interbank Offered Rate* (LIBOR) for the specific currency and maturity.

Why is BBA LIBOR important? For the US dollar in particular, LIBOR has become the primary benchmark interest rate for many short-term US dollar loans to corporations, including those

⁹ For example, if there are eight banks surveyed on a particular day, the two highest rates and the two lowest rates are dropped, and the average of the remaining four becomes the official LIBOR rate. If sixteen banks are surveyed, the four highest rates and the four lowest rates are dropped, and the average of the remaining eight becomes the official LIBOR rate.

¹⁰ Sometimes the bank that stamps the BA is different from the bank that actually lends the money to the corporation.

made in the US domestic market. The interest rate on these loans is quoted as a spread above or below LIBOR, such as ‘three-month LIBOR – 0.25%’ or ‘six-month LIBOR + 2.25%’. The size of the spread depends primarily on the credit quality of the customer borrowing the money. While most corporations can only borrow at a spread above LIBOR, some of the most creditworthy customers can obtain loans at rates below LIBOR.

US-dollar LIBOR is also the basis for settling many interest-rate futures contracts, including the most liquid contract of all, the three-month Eurodollar contract that trades on the Chicago Mercantile Exchange, as well as most over-the-counter interest-rate derivatives.

I.C.2.4 Money Market Securities

Money market securities are loans that have been structured so that they can be traded among investors in the secondary market with a wide variety of structures and characteristics. Money market securities are initially issued with terms of one year or less. They allow investors to place their excess cash in short-term instruments that, all else being equal, are less risky than securities with longer terms. They also allow investors to get a higher rate of return than they would from money sitting in a traditional bank account, while at the same time providing the issuers of money market securities with a relatively low-cost source of short-term funding.

While it is true that money market securities have lower risk than longer-dated bonds, market participants are exposed to risk at a number of levels, namely:

- *Interest-rate risk.* The risk that interest rates will rise (fall) and the price of the security will accordingly fall (rise). See Chapter I.B.2 for further discussion of the inverse relationship between interest rates and prices of securities.
- *Credit risk.* The risk that the issuer of the security will default on its obligations to repay interest, principal or both.
- *Liquidity risk.* The risk that an investor wishing to sell a security is not able to do so quickly without sacrificing price.

Money market securities often trade in denominations that are too large for individual investors. *Cash management trusts* or *money market mutual funds* have become popular vehicles that enable the small investor to participate in these markets. Such funds pool the resources of many investors and can trade in money market securities on their behalf, managing the risks in accordance with the fund’s trust deed.

I.C.2.4.1 Treasury Bills

Treasury bills (also known as *T-bills*) are short-term securities issued by governments, normally national governments, often using auction mechanisms, as part of their liquidity management operations. Credit risk is very low or effectively non-existent, depending on the credit standing of the issuing government. Investors can purchase T-bills on either the primary or the secondary market; that is, either at auction or from a securities dealer. Liquidity is excellent in the secondary market.

T-bills do not explicitly pay interest. Instead, they are sold to investors for less than their face value; when they mature, they are repaid at their face value. The difference between the issue price and the face value represents the return on the investment for the purchaser. Securities that pay interest in this fashion are known as discount instruments because they are issued and traded at a discount (that is, a lower price) to their face value.

Because of different market conventions, the quoted yield on some government T-bills is not directly comparable to the quoted yield on others. For example, Canadian T-bill yields are known as *bond equivalent* or *money market* yields, while US T-bill yields are known as *bank discount* yields.

The following equation is used to calculate the (bond equivalent) yield of a T-bill from its price:

$$Y = \frac{FV - P}{P} \times \frac{ADC}{DTM}, \quad (\text{I.C.2.2})$$

where: Y = the yield

FV = the face value of the T-bill

P = the price of the T-bill

DTM = the number of days until maturity

ADC = the denominator of the relevant day count convention.

Alternatively, if we know the yield we can calculate its price. Rearranging (I.C.2.2) we have:

$$P = \frac{FV}{1 + \left(Y \times \frac{DTM}{ADC} \right)}. \quad (\text{I.C.2.3})$$

Example I.C.2.4

If a 90-day Government of Canada Treasury bill with a face value of \$1 million is offered for a price of \$990,000, the yield on the T-bill is:

$$\frac{\$1 \text{ million} - \$990,000}{\$990,000} \times \frac{365}{90} = 0.040965 = 4.0965\%.$$

Hence a 90-day Canadian T-bill with a face value of \$1 million offered at a yield of 4.0965% has a price calculated as:

$$P = \frac{\$1 \text{ million}}{1 + \left(0.040965 \times \frac{90}{365}\right)} = \$990,000.$$

I.C.2.4.2 Commercial Paper

Corporations in need of short-term financing usually borrow money from one or more banks. For large corporations with good credit ratings, an alternative to bank lending is the commercial paper market. Commercial paper is similar to T-bills. It does not pay explicit interest: it is issued at a discount to its face value and it matures at face value. Commercial paper is often backed by the liquidity of a bank line of credit. The existence of the line of credit, however, does not mean a guarantee of repayment to the investor. The yields on commercial paper (and other money market securities) will generally be higher than those on T-bills because of the relatively greater credit risk and liquidity risk.

Promissory notes are very similar to commercial paper. They are bills issued by a borrower without the credit support of a financial institution (see the following discussion of bankers' acceptances for comparison).

I.C.2.4.3 Bankers' Acceptances

A banker's acceptance (BA) is short-term debt that is issued by a corporation and guaranteed by a bank. BAs may be used to facilitate the purchase and sale of goods, either domestically or internationally, or to borrow money for any purpose. BAs trade on the credit quality of the accepting bank, not that of the originating corporation. We show by example how a banker's acceptance is created.

Example I.C.2.5

Suppose that Joe's Furniture Emporium (JFE) has a credit facility with Scotiabank that includes the option of issuing \$10 million of BAs at an interest rate equal to Scotiabank's 3-month BA rate plus 150 basis points. (One hundred basis points equals one percentage point.) BAs are marketable securities, and Scotiabank's BA rate represents the rate at which Scotiabank is willing to buy BAs in the marketplace.

JFE decides that it needs \$10 million in financing for three months and so it informs Scotiabank that it will draw down on its BA facility. If Scotiabank's three-month BA rate is currently 5%,

JFE will issue what is, in effect, a marketable IOU to Scotiabank to pay it \$10 million in three months' time. The IOU is then accepted by the bank, hence the name.

Once Scotiabank accepts the IOU from JFE, a banker's acceptance is created. Scotiabank will lend JFE the discounted value of \$10 million based on Scotiabank's BA rate of 5%, and at maturity, JFE will pay the face value to Scotiabank. The actual amount that JFE borrows from Scotiabank is calculated using equation (I.C.2.3), based on Scotiabank's BA rate and the number of days in the three-month period. If there are 91 days in the three-month period, JFE will receive

$$P = \frac{\$10 \text{ million}}{1 + (0.05 \times \frac{91}{365})} = \frac{\$10 \text{ million}}{1.012465753} = \$9,876,877.$$

The extra 150 basis points that JFE pays to Scotiabank represents the stamping fee, which is Scotiabank's compensation for effectively guaranteeing JFE's debt. JFE must pay the stamping fee to Scotiabank in advance. If we substitute the face value for the principal and the stamping fee (in decimal terms) for the interest rate, equation (I.C.2.1) can be used to calculate the stamping fee on a Canadian BA. So, JFE will immediately pay Scotiabank a stamping fee of

$$\$10 \text{ million} \times 0.015 \times \frac{91}{365} = \$37,397.$$

The net effect is that JFE has borrowed \$9,839,480 (\$9,876,877 – \$37,397) and will repay \$10 million at maturity.¹⁰

Using (I.C.2.1) to calculate the dollar value of the stamping fee results in an effective borrowing cost (on a bond-equivalent basis) that is slightly higher than if we simply added the stamping fee to Scotiabank's BA rate. This is because the value of the stamping fee is based on the face value of the BA rather than the amount lent. To calculate JFE's effective borrowing cost, use (I.C.2.2), substituting the amount borrowed for the 'price':

$$\frac{\$10 \text{ million} - \$9,839,480}{\$9,839,480} \times \frac{365}{91} = 0.06543 = 6.54\%.$$

The payment of the stamping fee has the effect of increasing JFE's cost of funds to 6.54% on a bond equivalent basis.

At this point, Scotiabank owns the BA. It may choose to hold it as part of its securities portfolio, or sell it to another market participant, such as a pension fund, mutual fund, or another bank that

wants to invest in a liquid, short-term money market security. When the BA matures, 91 days after it is issued, Scotiabank will pay the holder of the BA its \$10 million face value, which Scotiabank will in turn receive from JFE.

I.C.2.4.4 Certificates of Deposit

A certificate of deposit is a time deposit with a bank which can be traded. They generally have maturity of less than three months. They can be thought of as bank bills where the issuer is the bank. Not surprisingly, they pay the same yield as BAs. They are often used by banks for the purpose of asset/liability management.

I.C.2.5 Summary

Money markets provide lenders and borrowers with a great deal of flexibility in terms of borrowing and deposits. As a result, they tend to be very active markets as well. Below are some key points to remember:

- The key characteristics of money market instruments are their term, principal, interest rate, marketability, security and call or put features.
- Deposits come in three basic forms: demand deposits, notice deposits, and fixed-term deposits.
- Demand deposits are basically chequing accounts and generally do not pay interest.
- Notice deposits are savings accounts that generally pay a low, floating rate of interest that changes infrequently.
- Very large corporations can borrow large sums of money from a syndicate of banks. The loans can be structured as credit facilities, or may be one-off fixed- or floating-rate loans.
- A repurchase agreement, or repo, is a loan in which a borrower sells a security to a lender at one price and agrees to buy back the security on a future date at a higher price. The loan is collateralised by the security itself.
- The Eurocurrency market is the largest segment of the international market for deposits and loans. It is the market for fixed-rate term deposits and loans in a currency other than the local currency of the bank branch that is accepting the deposit or extending the loan.

- LIBOR stands for the London Interbank Offered Rate and represents an average rate on Eurocurrency deposits offered by eight large banks in the London market.
- LIBOR serves as an important benchmark for lending rates, both in the domestic market and the international market. LIBOR is also used as a reference rate for many interest-rate derivatives.
- Money market securities are initially issued with terms of one year or less. They generally do not pay interest, but trade in the secondary market at a price that is less than their face value. At maturity the issuer repays the face value, and the difference between the face value and the issue or purchase price is the interest on the loan.
- Treasury bills are issued by governments, primarily national governments.
- Commercial paper is issued by large corporations.
- Corporations issue bankers' acceptances (BAs) through a bank that guarantees the issue. BAs trade on the credit quality of the guarantor bank rather than the originating corporation. The guarantor bank charges the corporation a stamping fee to provide this guarantee.

I.C.3 Bond Markets

Moorad Choudhry, Lionel Martellini and Philippe Priaulet¹

This chapter describes the operation of bond markets. The focus is on market participants (Section I.C.3.2), the characteristics of bonds according to the type of issuer (Section I.C.3.3), the conventions and practices of bond markets (Section I.C.3.4) and the role of credit risk in bond markets (Section I.C.3.5). Section I.C.3.1 introduces the topic with a discussion of some important trends that have impacted the bond market.

I.C.3.1 Introduction

In most countries government expenditure exceeds the level of government income received through taxation. This shortfall is met by government borrowing, and bonds are issued to finance the government's debt. The core of any domestic capital market is usually the government bond market, which also forms the benchmark for all other borrowing. Government agencies also issue bonds, as do local governments or municipalities. Often (but not always) these bonds are virtually as secure as government bonds. Corporate borrowers issue bonds both to raise finance for major projects and also to cover ongoing and operational expenses. Corporate finance is a mixture of debt and equity, and a specific capital project will often be financed by a mixture of both.

The debt capital markets exist because of the financing requirements of governments and corporates. The source of capital is varied, but the total supply of funds in a market is made up of personal or household savings, business savings and increases in the overall money supply. However, the requirements of savers and borrowers differ significantly, in that savers have a short-term investment horizon while borrowers prefer to take a longer-term view. The 'constitutional weakness' of what would otherwise be unintermediated financial markets led, from an early stage, to the development of financial intermediaries.

The world bond market has increased in size more than 15 times in the last thirty years. As at the end of 2002 outstanding volume stood at over \$21 trillion. Table I.C.3.1 shows that the United States constitutes almost 50% of the world's bond market.

¹ Lionel Martellini is a Professor of Finance at EDHEC Graduate School of Business, and the Scientific Director of EDHEC Risk and Asset Management Research Center (www.edhec-risk.com). Moorad Choudhry is Visiting Professor, Department of Economics, Finance and International Business, London Metropolitan University. Philippe Priaulet is a Fixed-Income Strategist in charge of derivatives strategies for HSBC, and also an Associate Professor in the Department of Mathematics of the University of Evry Val d'Essonne and a lecturer at ENSAE.

Table I.C.3.1: Major government bond markets, December 2002

<i>Country</i>	Nominal value (\$ billion)	Percentage of World Market
<i>United States</i>	5,490	48.5
<i>Japan</i>	2,980	26.3
<i>Germany</i>	1,236	10.9
<i>France</i>	513	4.5
<i>Canada</i>	335	3.0
<i>United Kingdom</i>	331	2.9
<i>Netherlands</i>	253	2.2
<i>Australia</i>	82	0.7
<i>Denmark</i>	72	0.6
<i>Switzerland</i>	37	0.3
<i>Total</i>	11,329	100

Source: IFC 2003.

The origin of the spectacular increase in the size of global financial markets was the rise in oil prices in the early 1970s. Higher oil prices stimulated the development of a sophisticated international banking system, as they resulted in large capital inflows to developed country banks from the oil-producing countries. A significant proportion of these capital flows were placed in *Eurodollar* deposits in major banks. The growing trade deficit and level of public borrowing in the United States also contributed. The last twenty years have seen tremendous growth in capital markets' volumes and trading. As capital controls were eased and exchange rates moved from fixed to floating, domestic capital markets became internationalised. Growth was assisted by the rapid advance in information technology and the widespread use of financial engineering techniques. Today we would think nothing of dealing in virtually any liquid currency bond in financial centres around the world, often at the touch of a button. Global bond issues, underwritten by the subsidiaries of the same banks, are commonplace. The ease with which transactions can be undertaken has also contributed to a very competitive market in liquid currency assets.

I.C.3.2 The Players

A wide range of participants are involved in the bond markets. We can group them broadly into borrowers and investors, plus the institutions and individuals who are part of the business of bond trading. Borrowers access the bond markets as part of their financing requirements; hence, borrowers can include sovereign governments, local authorities, public sector organisations and

corporations. Virtually all businesses operate with a financing structure that is a mixture of debt and equity finance, and debt finance almost invariably contains a form of bond finance.

I.C.3.2.1 Intermediaries and Banks

In its simplest form a financial intermediary is a broker or agent. Today we would classify the broker as someone who acts on behalf of the borrower or lender, buying or selling a bond as instructed. However, intermediaries originally acted between borrowers and lenders in placing funds as required. A broker would not simply on-lend funds that had been placed with it, but would accept deposits and make loans as required by its customers. This resulted in the first banks.

A *retail bank* deals mainly with the personal financial sector and small businesses, and in addition to loans and deposits also provides cash transmission services. A retail bank is required to maintain a minimum cash reserve, to meet potential withdrawals, but the remainder of its deposit base can be used to make loans. This does not mean that the total size of its loan book is restricted to what it has taken in deposits: loans can also be funded in the wholesale market.

An *investment bank* will deal with governments, corporates and institutional investors. Investment banks perform an agency role for their customers and are the primary vehicle through which a corporate will borrow funds in the bond markets. This is part of the bank's corporate finance function. It will also act as wholesaler in the bond markets, a function known as *market making*. The bond issuing function of an investment bank, by which the bank will issue bonds on behalf of a customer and pass the funds raised to this customer, is known as *origination*. Investment banks will also carry out a range of other functions for institutional customers, including export finance, corporate advisory services and fund management. Other financial intermediaries will trade not on behalf of clients but for their own book. These include arbitrageurs and speculators. Usually such market participants form part of investment banks.

I.C.3.2.2 Institutional Investors

We can group the main types of institutional investors according to the time horizon of their investment activity:

- *Short-term institutional investors*. These include banks and building societies, money market fund managers, central banks and the treasury desks of some types of corporates. Such bodies are driven by short-term investment views, often subject to close guidelines. Banks will have an additional requirement to maintain liquidity, often in fulfilment of regulatory authority rules, by holding a proportion of their assets in the form of short-term instruments that are easy to trade.

- *Long-term institutional investors.* Typically these types of investors include pension funds and life assurance companies. Their investment horizon is long-term, reflecting the nature of their liabilities. Often they will seek to match these liabilities by holding long-dated bonds.
- *Mixed horizon institutional investors.* This is possibly the largest category of investors and will include general insurance companies and most corporate bodies. Like banks and financial sector companies, they are also very active in the primary market, issuing bonds to finance their operations.

I.C.3.2.3 Market Professionals

These players include the banks and specialist financial intermediaries mentioned above, firms that one would not automatically classify as ‘investors’ although they will also have an investment objective. Their time horizon will range from one day to the very long term. They include:

- proprietary trading desks of investment banks;
- bond market makers in securities houses and banks providing a service to their customers;
- inter-dealer brokers that provide an anonymous broking facility.

Proprietary traders will actively position themselves in the market in order to gain trading profit, for example in response to their view on where they think interest rate levels are headed. These participants will trade direct with other market professionals and investors, or via brokers.

Market makers or ‘traders’ (also called ‘dealers’ in the United States) are wholesalers in the bond markets; they make two-way prices in selected bonds. Firms will not necessarily be active market makers in all types of bonds; smaller firms often specialise in certain sectors. In a two-way quote the *bid price* is the price at which the market maker will buy stock, so it is the price the investor will receive when selling stock. The *offer price* or ask price is the price at which investors can buy stock from the market maker. As one might expect, the bid price is always higher than the offer price, and it is this spread that represents the theoretical profit to the market maker. The bid–offer spread set by the market maker is determined by several factors, including supply and demand, and liquidity considerations for that particular stock, the trader’s view on market direction and volatility, as well as that of the stock itself and the presence of any market intelligence. A large bid–offer spread reflects low liquidity in the stock, as well as low demand.

To facilitate a liquid market there also exist *inter-dealer brokers* (IDBs). These provide an anonymous broking facility so that market makers can trade in size at the keenest prices. Generally IDBs will post prices on their screens that have been provided by market makers on a no-names basis. The screens are available to other market makers (and in some markets to other

participants as well). At any time IDB screen prices represent the latest market price and bid–offer spread. IDBs exist in government, agency, corporate and Eurobond markets.

I.C.3.3 Bonds by Issuers

This section describes the main classes of bonds by type of borrower. On the public side we distinguish between *sovereign bonds* issued by national governments, *agency bonds* issued by public bodies and *municipal bonds* issued by local governments. On the private side we have the *corporate bonds* issued by corporations, and we further distinguish between *domestic* and *foreign bonds*, and *international bonds*, the latter constituting the large class of *Eurobonds*. Here we discuss the special characteristics of each of these types of bond.

I.C.3.3.1 Government Bonds

The four major government bond issuers in the world are the euro-area countries, Japan, the United States and, to a lesser extent, the United Kingdom. Table I.C.3.2 gives a country percentage breakdown of the JP Morgan Global Government Bond Index, which is a benchmark index for developed government debt markets.

Table I.C.3.3 compares the features of the world’s most important government bond markets. Note the minor variations in market practice with regard to the frequency of coupons, the day-count basis, benchmark bonds, etc. Most government bonds are issued by a standard auction process, where the price is gradually reduced until it meets a bid. The sale price varies for each successful bidder, depending on the bid price. Others use the so-called Dutch auction system. Under this system the securities are allocated to bidders starting with the highest bid. The price at which the final allocation is made becomes the price at which *all* securities are sold.

Table I.C.3.4 shows the country yield curves at the time of writing, and a subset of them are graphed in Figure I.C.3.1. Note the variability in yield curves between countries reflecting their varying economic conditions and risk profiles. While most have an upward-sloping (or normal) yield curve, two of the yield curves (UK and Australia) are quite flat. Discussion of the various theories explaining the shape of the yield curve can be found in Chapter I.A.6.

Table I.C.3.2: JP Morgan Global Government Bond Index (October 2001)

Market	USD Value	Weight in Index	Daily Yield	Macaulay Duration	Convexity	Remaining Maturity
Global Index	248.262	100%	4.67	5.888	63.154	8.153
Australia	312.251	0.44%	5.088	4.678	30.824	5.855
Belgium	226.492	3.02%	4.675	5.591	49.436	7.339
Canada	297.4	2.64%	5.192	5.842	67.935	9.22
Denmark	262.114	1.27%	4.606	4.27	33.061	5.578
France	243.236	8.75%	4.583	5.307	51.392	7.335
Germany	181.024	9.26%	4.565	5.318	56.072	7.445
Italy	266.716	8.18%	4.922	5.609	66.359	8.586
Japan	212.269	27.75%	0.971	5.703	48.306	6.197
Netherlands	192.365	2.34%	4.65	5.427	54.334	7.494
Spain	255.276	3.44%	4.734	5.414	52.995	7.364
Sweden	218.143	0.84%	4.943	4.642	32.339	5.665
United Kingdom	291.253	5.24%	4.759	7.534	100.694	11.598
United States	304.682	26.81%	4.867	6.504	82.568	10.266
New Zealand	126.373		6.003	4.313	28.153	5.396
Ireland	262.434		4.756	6.768	61.241	8.31
Finland	130.401		4.407	4.504	29.214	5.338
Portugal	142.856		4.646	5.572	42.416	6.705
South Africa	175.266		10.385	5.432	44.948	9.634

© 2001 J.P. Morgan & Co. Incorporated

Table I.C.3.3: Government bond markets: characteristics of selected countries

Source: Choudhry (2004)

	Credit rating	Maturity range	Dealing	Benchmark bonds	Issuance	Coupon and day-count basis
Australia	AAA	2–15 years	OTC Dealer network	5, 10 years	Auction	Semi-annual, act/act
Canada	AAA	2–30 years	OTC Dealer network	3, 5, 10 years	Auction, subscription	Semi-annual, act/act
France	AAA	BTAN: 1–7 years OAT: 10–30 years	OTC Dealer network. listed on Paris Stock Exchange	Bonds BTAN: 2 and 5 year OAT: 10 and 30 years	Dutch auction	BTAN: Semi-annual, act/act OAT: Annual, act/act
Germany	AAA	OBL: 2, 5 years BUND: 10, 30 years	OTC Dealer network. on Stock Exchange	Listed The most recent issue	Combination of Dutch auction and proportion of each issue allocated on fixed basis to institutions	Annual, act/act
South Africa	A	2–30 years	OTC Dealer network. on Johannesburg SE	Listed 2, 7, 10 and 20 years	Auction	Semi-annual, act/365
Singapore	AAA	2–15 years	OTC Dealer network	1, 5, 10 and 15 years	Auction	Semi-annual, act/act
Taiwan	AA-	2–30 years	OTC Dealer network	2, 5, 10, 20 and 30 years	Auction	Annual, act/act
United Kingdom	AAA	2–35 years	OTC Dealer network	5, 10, 30 years	Auction, subsequent issue by ‘tap’ subscription	Semi-annual, act/act
United States	AAA	2–20 years	OTC Dealer network	2, 5, 10 years	Auction	Semi-annual, act/act

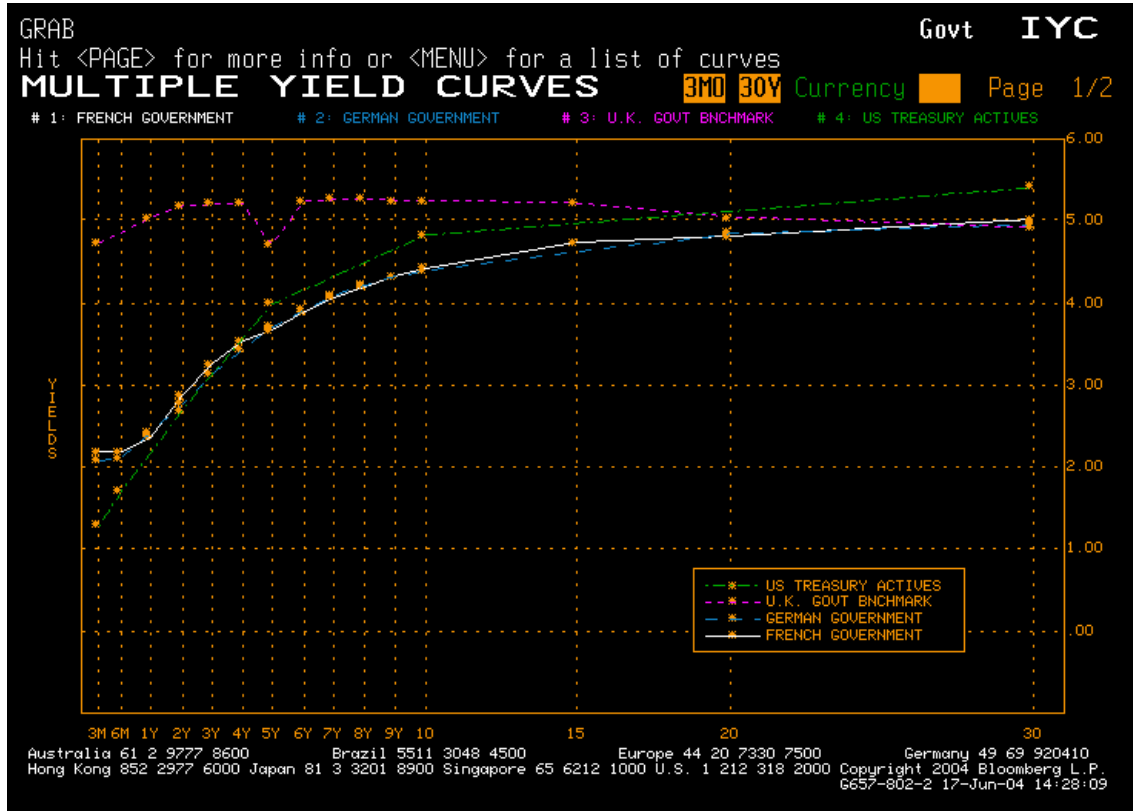
Table I.C.3.4: Country yield curves (as of 21 June 2004)

Yield source: Bloomberg L.P.

Years to maturity	Australia	Canada	France	Germany	South Africa	Singapore	Taiwan	United Kingdom	United States
1	5.2154	2.749	2.3291	2.3747	4.71	0.937		4.9599	
2	5.2259	3.394	2.8375	2.7184	9.351	1.157	1.449	5.1033	2.7034
3	5.3451	2.6339	3.1998	3.0609				5.1633	3.1129
4	5.482	3.937	3.4966	3.3998	10.025			5.1685	
5	5.5761	4.2684	3.7222	3.6425		2.3085	2.3243	4.6946	3.9406
7	5.735	4.0704	4.014	4.0465		2.9583		5.2042	
10	5.8888	4.9984	4.395	4.3708	10.468	3.3355	3.0608	5.1863	4.8135
15	5.941		4.709			3.8989	3.1635	5.1504	
20		5.2426	4.776	4.8365	9.605		3.3507	4.9885	
30		5.4447	4.98	4.9481			3.5571	4.8596	5.3878

Figure I.C.3.1: Bloomberg screen IYC showing yield curves for US, UK, French and German government bond markets, 17 June 2004

© Bloomberg L.P. Used with permission. All rights reserved. Visit www.bloomberg.com



In the US case, government securities are issued by the US Department of the Treasury and backed by the full faith and credit of the US government. These are called ‘Treasury securities’. The Treasury market is the most active market in the world, thanks to the large volume of total debt and the large size of any single issue. The amount of outstanding marketable US Treasury securities is huge, with a value of \$3.4 trillion as of December 2003. The Treasury market is the most liquid debt market, that is, the one where pricing and trading are most efficient. The bid–offer spread is far lower than in the rest of the bond market. Recently issued Treasury securities are referred to as *on-the-run* securities, as opposed to *off-the-run* securities, which are old issued securities. Special mention must be made of benchmark securities, which are recognized as market indicators. There typically exists one such security on each of the following yield curve points: 2 years, 5 years, 10 years and 30 years. As they are over-liquid they trade richer than all of their direct neighbours.

Example I.C.3.1

On 7 December 2001, the 5-year US Treasury benchmark bond had a coupon of 3.5% and a maturity date of 15 November 2006. It had been issued on 15 November 2001. In contrast, a 5-year off-the-run US T-bond had a coupon of 7% and a maturity date of 15 July 2006. Its issuance date was 15 July 1996. Note the difference of coupon level between the two. There are two reasons for that: first, the 5-year off-the-run T-bond was originally a 10-year T-bond. Its coupon reflected the level of 10-year yields at that time. Second, the level of the US government yield curve on 15 July 1996 was at least 200 basis points over the level of the US government yield curve on 15 November 2001. Furthermore, on 7 December 2001, the yield of the off-the-run bond was 4.48% as opposed to 4.45% for the benchmark bond, which illustrates the relative richness of the latter. Government bonds are traded on major exchanges as well as *over the counter*.² The New York Stock Exchange had over 660 government issues listed on it at the end of 2003, with a total par value of \$3.1 billion.

I.C.3.3.2 US Agency Bonds

These are issued by different organizations, seven of which dominate the US market in terms of outstanding debt: the Federal National Mortgage Association (Fannie Mae), the Federal Home Loan Bank System (FHLBS), the Federal Home Loan Mortgage Corporation (Freddie Mac), the Farm Credit System (FCS), the Student Loan Marketing Association (Sallie Mae), the Resolution Funding Corporation (RefCorp) and the Tennessee Valley Authority (TVA). Agencies have at least two common features:

- *They were created to fulfil a public purpose.* For example in the USA, Fannie Mae and Freddie Mac aim to provide liquidity for the residential mortgage market. The FCS aims to

² Generally OTC refers to trades that are not carried out on an exchange but directly between the counterparties.

support agricultural and rural lending. RefCorp aims to provide financing to resolve thrift crisis.

- *The debt is not necessarily guaranteed by the government.* Hence it contains a credit premium. In fact in the USA, there are a few federally related institution securities, such as the Government National Mortgage Association (GNMA), and these *are* generally backed by the full faith and credit of the US government. There is no credit risk, but since they are relatively small issues they contain a liquidity premium.

Agencies are differently organised. For instance, Fannie Mae, Freddie Mac and Sallie Mae are owned by private-sector shareholders, the FCS and the FHLBS are cooperatives owned by the members and borrowers. One sizeable agency, the Tennessee Valley Authority, is owned by the US government.

I.C.3.3.3 Municipal Bonds

Municipal securities constitute the municipal market, that is, the market where state and local governments – counties, special districts, cities and towns – raise funds in order to finance projects for the public good such as schools, highways, hospitals, bridges and airports. Typically, bonds issued in this sector are exempt from federal income taxes, so this sector is referred to as the *tax-exempt sector*. There are two generic types of municipal bonds: *general obligation bonds* and *revenue bonds*. The former have principal and interest secured by the full faith and credit of the issuer and are usually supported by either the issuer's unlimited or limited taxing power. The latter have principal and interest secured by the revenues generated by the operating projects financed with the proceeds of the bond issue. Many of these bonds are issued by special authorities created for the purpose.

I.C.3.3.4 Corporate Bonds

Corporate bonds are issued by entities belonging to the private sector. They represent what market participants call the credit market. In the corporate markets, bond issues usually have a stated term to maturity, although the term is often not fixed because of the addition of call or put features. The convention is for most corporate issues to be medium- or long-dated, and rarely to have a term greater than 20 years. In the US market prior to the Second World War it was once common for companies to issue bonds with maturities of 100 years or more, but this is now quite rare. Only the highest-rated companies find it possible to issue bonds with terms to maturity greater than 30 years; during the 1990s such companies included Coca-Cola, Disney and British Gas.

Investors prefer to hold bonds with relatively short maturities because of the greater price volatility experienced in the markets since the 1970s, when high inflation and high interest rates were common. A shorter-dated bond has lower interest-rate risk and price volatility than a longer-dated bond. There is thus a conflict between investors, whose wish is to hold bonds of shorter maturities, and borrowers, who would like to fix their borrowing for as long a period as possible. Although certain institutional investors such as pension fund managers have an interest in holding 30-year bonds, it would be difficult for all but the largest, best-rated companies, to issue debt with a maturity greater than this. Highly rated corporate borrowers are often able to issue bonds without indicating specifically how they will be redeemed. By implication, maturity proceeds will be financed out of the company's general operations or by the issue of another bond. However, borrowers with low ratings may make specific provisions for paying off a bond issue on its maturity date, to make their debt issue more palatable to investors. For instance, a ring-fenced sum of cash (called the *sinking fund*) may be put aside to form the proceeds used in the repayment of a fixed-term bond. A proportion of a bond issue is redeemed every year until the final year when the remaining outstanding amount is repaid. In most cases the issuer will pass the correct cash proceeds to the bond's trustee, who will use a lottery method to recall bonds representing the proportion of the total nominal value outstanding that is being repaid. The trustee usually publishes the serial numbers of bonds that are being recalled in a newspaper such as the *Wall Street Journal* or the *Financial Times*. The price at which bonds are redeemed by a sinking fund is usually par. If a bond has been issued above par, the sinking fund may retire the bonds at the issue price and gradually decrease this each year until it reaches par.³ Sinking funds reduce the credit risk applying to a bond issue, because they indicate to investors that provision has been made to repay the debt. However, there is a risk associated with them, in that at the time bonds are paid off they may be trading above par due to a decline in market interest rates. In this case investors will suffer a loss if it is their holding that is redeemed.

Bonds that are secured through a charge on fixed assets such as property or plant often have certain clauses in their offer documents that state that the issuer cannot dispose of the assets without making provision for redemption of the bonds, as this would weaken the collateral backing for the bond. These clauses are known as *release-of-property* and *substitution-of-property* clauses. Under these clauses, if property or plant is disposed, the issuer must use the proceeds (or part of the proceeds) to redeem bonds that are secured by the disposed assets. The price at which the bonds are retired under this provision is usually par, although a special redemption price other than par may be specified in the repayment clause.

³ Another method by which bonds are repaid is that the issuer will purchase the required nominal value of the bonds in the open market; these are then delivered to the trustee, who cancels them.

A large number of corporate bonds, especially in the US market, have a *call provision*. Borrowers prefer this as it enables them to refinance debt at cheaper rates when market interest rates have fallen significantly below their level at the time of the bond issue. A call provision is a negative feature for investors, as bonds are only paid off if their price has risen above par. Although a call feature indicates an issuer's interest in paying off the bond, because they are not attractive for investors, callable bonds pay a higher yield than non-callable bonds of the same credit quality.

In general, callable bonds are not callable for the first 5–10 years of their life, a feature that grants an element of protection for investors. Thereafter a bond is usually callable on set dates up to the final maturity date. In the US market another restriction is that of refunding redemption. This prohibits repayment of bonds within a set period after issue with funds obtained at a lower interest rate or through issue of bonds that rank with or ahead of the bond being redeemed. A bond with refunding protection during the first 5–10 years of its life is not as attractive as a bond with absolute call protection. Bonds that are called are usually called at par, although it is common also for bonds to have a call schedule that states that they are redeemable at specified prices above par during the call period.

Corporate bonds are traded on exchanges and OTC. Outstanding volume as at the end of 2003 was \$8.1 trillion (see Choudhry, 2004). The corporate bond market varies in liquidity, depending on the currency and type of issuer of any particular bond. As in the case of sovereign bonds, liquidity is greater for recent issues. But corporate bonds in general are far less liquid than government bonds: they bear higher bid–ask spreads.

Example I.C.3.2

On 10 December 2001, the bid–ask price spread for the T-bond 6% 15 August 2009 amounted to 1.5 cents, whereas for the Ford corporate bond 7.375% 28 October 2009 it amounted to 60 cents. The pricing source that is used is the Bloomberg Generic Value, that is, the average of the prices of the most active contributors. It is a market consensus price.

One of the most liquid corporate bond types is the Eurobond, which is an international bond issued and traded across national boundaries. This is discussed below.

I.C.3.3.5 Eurobonds (International Bonds)

In any market there is a primary distinction between *domestic* bonds and other bonds. Domestic bonds are issued by borrowers domiciled in the country of issue, and in the currency of the country of issue. Generally they trade only in their original market. A *Eurobond* is issued across national boundaries and can be in any currency, which is why they are also sometimes called

international bonds. In fact, it is now more common for Eurobonds to be referred to as international bonds, to avoid confusion with ‘euro bonds’, which are bonds denominated in *euros*, the currency of 12 countries of the European Union (EU). As an issue of international bonds is not restricted in terms of currency or country, the borrower is not restricted as to its nationality either. There are also *foreign* bonds, which are domestic bonds issued by foreign borrowers. An example of a foreign bond is a *Bulldog*, which is a sterling bond issued for trading in the UK market by a foreign borrower. The equivalent foreign bonds in other countries include *Yankee* bonds (USA), *Samurai* bonds (Japan), *Alpine* bonds (Switzerland) and *Matador* bonds (Spain). There are detailed differences between these bonds, for example in the frequency of interest payments that each one makes and the way the interest payment is calculated. Some bonds, such as domestic bonds, pay their interest *net*, which means net of a withholding tax such as income tax. Other bonds, including Eurobonds, make *gross* interest payments.

Nowhere has the increasing integration and globalisation of the world’s capital markets been more evident than in the Eurobond market. It is an important source of funds for many banks and corporates, not to mention central governments. The Eurobond market continues to develop new structures in response to the varying demands and requirements of specific groups of investors. Often the Eurobond market is the only opening for certain types of government and corporate finance. Investors also look to the Eurobond market due to constraints in their domestic market, and Euro securities have been designed to reproduce the features of instruments that certain investors may be prohibited from investing in domestically. Other instruments are designed for investors in order to provide tax advantages. The traditional image of the Eurobond investor, the so-called ‘Belgian dentist’, has changed and the investor base is both varied and geographically dispersed.

I.C.3.4 The Markets

A distinction is made between financial instruments of up to one year’s maturity and instruments of over one year’s maturity. Short-term instruments make up the *money market* while all other instruments are deemed to be part of the *capital market*. There is also a distinction made between the *primary market* and the *secondary market*. A new issue of bonds made by an investment bank on behalf of its client is made in the primary market. Such an issue can be a *public offer*, in which anyone can apply to buy the bonds, or a *private offer*, where the customers of the investment bank are offered the stock. The secondary market is the market in which existing bonds are subsequently traded.

Bond markets are regulated as part of the overall financial system. In most countries there is an independent regulator responsible for overseeing both the structure of the market and the bonafides of market participants. For instance, the US market regulator is the Securities and Exchange Commission (SEC). The UK regulator, the Financial Services Authority (FSA), is responsible for regulating both wholesale and retail markets; for example, it reviews the capital requirements for commercial and investment banks, and it is also responsible for regulating the retail mortgage market. Money markets are usually overseen by the country's central bank – for example, the Federal Reserve manages the daily money supply in the USA, while the Bank of England provides liquidity to the market by means of its daily money market repo operation.

I.C.3.4.1 The Government Bond Market

Government bonds are traded on the following four markets: in addition to the primary and secondary markets, we have the *when-issued market* and the *repo market*.

- The primary market: newly issued securities are first sold through an auction, which is conducted on a competitive bid basis. The auction process happens between the government and primary/non-primary dealers according to regular cycles for securities with specific maturities.⁴
- The secondary market: here a group of government securities dealers offer continuous bid and ask prices on specific outstanding government bonds. This is an OTC market.
- The when-issued market: here securities are traded on a *forward* basis before they are issued by the government.
- The repo market: in this market securities are used as collateral for loans. A distinction must be made between the *general-collateral* repo rate (GC) and the *special* repo rate. The GC repo rate applies to the major part of government securities. Special repo rates are specific repo rates. They typically concern on-the-run and cheapest-to-deliver securities, which are very expensive. This is the reason why special repo rates are at a level below the GC repo rate. Indeed, as these securities are very much in demand, the borrowers of these securities on the repo market receive a relatively lower repo rate compared to normal government securities (see Chapter I.C.2).

The bonds issued by regional governments and certain public sector bodies, such as national power and telecommunications utilities, are usually included as 'government' debt, as they are almost always covered by an explicit or implicit government guarantee. All other categories of

⁴ The US auction cycles are as follows: two-year notes are auctioned every month and settle on the 15th. Five-year notes are auctioned quarterly, in February, May, August and November of each year, and settle at the end of the month. Ten-year notes are auctioned quarterly, in February, May, August and November of each year, and settle on the 15th of the month. Thirty-year bonds are auctioned semi-annually: in February and August of each year, and settle on the 15th of the month. Auctions are announced by the Treasury one week in advance, the issuing date being set one to five days after the auction.

borrower are therefore deemed to be ‘corporate’ borrowers. Generally the term ‘corporate markets’ is used to cover bonds issued by non-government borrowers.

I.C.3.4.2 The Corporate Bond Market

In the context of a historically low level of interest rates, linked to a decreasing trend in inflation as well as in budget deficits, the corporate bond market is rapidly developing and growing. This strong tendency affects both the supply and the demand. While corporate supply is expanding, in relation to bank disintermediation, corporate demand is rising as more and more investors accustomed to dealing with only government bonds are including corporate bonds in their portfolios so as to capture spread and generate performance.

I.C.3.4.2.1 The market by country and sector

Within the four major bond markets in the world, the US dollar (USD) corporate market is the most mature, followed by the sterling (GBP) market and the euro (EUR) market, the growth of the latter being reinforced by the launching of the euro. The Japanese yen (JPY) market differentiates itself from the others, because of the credit crunch situation and economic difficulties it has been facing since the Asian crisis. The USD corporate bond market is the largest and most diversified: it is for instance more than twice as big as the Euro market, and low investment-grade ratings are much more represented (being over 80% of the index).

The corporate bond market can be divided into three main sectors: financial, industrial, utility. Apart from the USD market, the financial sector is over-represented. It is another proof of the maturity of the USD market, where the industrial sector massively uses the market channel in order to finance investment projects. It is also worth noting that the sector composition in the USD market is far more homogeneous than in the other markets. For example, the banking sector is systematically predominant in the GBP, EUR and JPY financial markets, while the telecommunication sector exceeds one third of the Euro industrial market. As a result, local credit portfolio diversification can be better achieved in the USD market than in the others.

I.C.3.4.2.2 Underwriting a new issue

The issue of corporate debt in the capital markets requires a primary market mechanism. The first requirement is a collection of merchant banks or investment banks that possess the necessary expertise. Investment banks provide advisory services on corporate finance as well as underwriting services, which is a guarantee to place an entire bond issue into the market in return for a fee. As part of the underwriting process the investment bank will either guarantee a minimum price for the bonds, or aim to place the paper at the best price available. The major

underwriting institutions in emerging economies are often branch offices of the major integrated global investment banks.

Small size bond issues may be underwritten by a single bank. It is common, however, for larger issues, or issues that are aimed at a cross-border investor base, to be underwritten by a syndicate of investment banks. This is a group of banks that collectively underwrite a bond issue, with each syndicate member being responsible for placing a proportion of the issue. The bank that originally won the mandate to place the paper invites other banks to join the syndicate. This bank is known as the *lead underwriter*, *lead manager* or *book-runner*. An issue is brought to the market simultaneously by all syndicate members, usually via the *fixed price re-offer* mechanism.⁵ This is designed to guard against some syndicate members in an offering selling stock at a discount in the *grey market*, to attract investors.⁶ This would force the lead manager to buy the bonds back if it wished to support the price. Under the fixed price re-offer method, price undercutting is not possible as all banks are obliged not to sell their bonds below the initial offer price that has been set for the issue. The fixed price usually is in place up to the first settlement date, after which the bond is free to trade in the secondary market.

I.C.3.4.3 The Eurobond Market

The key feature of a Eurobond is the way it is issued, internationally across borders and by an international underwriting syndicate. The method of issuing Eurobonds reflects the cross-border nature of the transaction and, unlike government markets where the auction is the primary issue method, Eurobonds are typically issued under a fixed price re-offer method (see Section I.C.3.4.2.2) or a *bought deal*.⁷

⁵ In a fixed price re-offer scheme the lead manager will form the syndicate, which will agree on a fixed issue price, a fixed commission and the distribution amongst themselves of the quantity of bonds they will take as part of the syndicate. The banks then re-offer the bonds that they have been allotted to the market, at the agreed price. This technique gives the lead manager greater control over an issue. It sets the price at which other underwriters in the syndicate can initially sell the bonds to investors. The fixed price re-offer mechanism is designed to prevent underwriters from selling the bonds back to the lead manager at a discount to the original issue price, that is, 'dumping' the bonds.

⁶ The grey market is a term used to describe trading in the bonds before they officially come to the market, mainly market makers selling the bond short to other market players or investors. Activity in the grey market serves as useful market intelligence to the lead manager, who can gauge the level of demand that exists in the market for the issue. A final decision on the offer price is of course not made until the actual issue date.

⁷ In a bought deal, a lead manager or a managing group approaches the issuer with a firm bid, specifying issue price, amount, coupon and yield. Only a few hours are allowed for the borrower to accept or reject the terms. If the bid is accepted, the lead manager purchases the entire bond issue from the borrower. The lead manager then has the option of selling part of the issue to other banks for distribution to investors, or doing so itself. In a volatile market the lead manager will probably parcel some of the issue to other banks for placement. However, it is at this time that the risk of banks dumping bonds on the secondary market is highest; in this respect lead managers will usually pre-place the bonds with institutional investors before the bid is made. The bought deal is focused primarily on institutional rather than private investors. As the syndicate process is not used, the bought deal requires a lead manager with sufficient capital and placement power to enable the entire issue to be placed.

The range of borrowers in the Euromarkets is very diverse. From virtually the inception of the market, borrowers representing corporates, sovereign and local governments, nationalised corporations, supranational institutions and financial institutions have raised finance in the international markets. The majority of borrowing has been by national governments, regional governments and public agencies of developed countries, although the Eurobond market is increasingly a source of finance for developing country governments and corporates.

Governments and institutions access the Euromarkets for a number of reasons. Under certain circumstances it is more advantageous for a borrower to raise funds outside its domestic market, due to the effects of tax or regulatory rules.⁸ The international markets are very competitive in terms of using intermediaries, so a borrower may well be able to raise cheaper funds in the international markets.

Other reasons why borrowers access Eurobond markets include:

- a desire to diversify sources of long-term funding. A bond issue is often placed with a wide range of institutional and private investors, rather than the more restricted investor base that may prevail in a domestic market. This gives the borrower access to a wider range of lenders, and for corporate borrowers this also enhances the international profile of the company.
- for both corporates and emerging country governments, the prestige associated with an issue of bonds in the international market.
- the flexibility of a Eurobond issue compared to a domestic bond issue or bank loan, illustrated by the different types of Eurobond instruments available.

Against this are balanced the potential downsides of a Eurobond issue, which include the following:

- for all but the largest and most creditworthy of borrowers, the rigid nature of the issue procedure becomes significant during times of interest-rate and exchange-rate volatility, reducing the funds available for borrowers;
- issuing debt in currencies other than those in which a company holds matching assets, or in which there are no prospects of earnings, exposes the issuer to foreign exchange risk.

Table I.C.3.5 shows some of the outstanding issues in the Eurobond market in 1999. The market remains an efficient and attractive market in which a company can raise finance for a wide range of maturities. The institutional investors include insurance companies, pension funds, investment trusts, commercial banks, and corporations – just as in domestic corporate bond markets. Other

⁸ There is no formal regulation of the Eurobond market as such, but each market participant will be subject to the regulation of its country regulator.

investors include central banks and government agencies; for example, the Kuwait Investment Office and the Saudi Arabian Monetary Agency both have large Eurobond holdings. In the UK, banks and securities houses are keen holders of Eurobonds issued by other financial institutions.

Table I.C.3.5: Selected euro-denominated Eurobond issues in 1999

Issuer	Rating	Coupon	Maturity	Volume (€m)	Launch spread (benchmark) bps
Pearson	Baa1/BBB+	4.625%	July 2004	400	82
Lafarge	A3/A	4.375%	July 2004	500	52
Mannesmann	A2/A	4.875%	September 2004	2500	75
Enron	Baa2/BBB+	4.375%	April 2005	400	90
Swissair	–	4.375%	June 2006	400	78
Renault	Baa2/BBB+	5.125%	July 2006	500	88
Continental Rubber		5.25%	July 2006	500	100
Yorkshire Water	A2/A+	5.25%	July 2006	500	75
British Steel	A3/A–	5.375%	August 2006	400	105
International Paper	A3/BBB+	5.375%	August 2006	250	105
Hammerson	Baa1/A	5%	July 2007	300	92
Mannesmann	A2/A	4.75%	May 2009	3000	70

Source: Westdeutsche Landesbank

I.C.3.4.4 Market Conventions

A particular market will apply one of five different methods to calculate accrued interest:

actual/365 Accrued = Coupon days/365

actual/360 Accrued = Coupon days/360

actual/actual Accrued = Coupon days/actual number of days in the interest period

30/360 See below

30E/360 See below

Table I.C.3.6: Selected bond market conventions

Market	Coupon frequency	Day count basis	Ex-dividend period
Australia	Semi-annual	actual/actual	Yes
Austria	Annual	actual/actual	No
Belgium	Annual	actual/actual	No
Canada	Semi-annual	actual/actual	No
Denmark	Annual	actual/actual	Yes
Eurobonds	Annual	30/360	No
France	Annual	actual/actual	No
Germany	Annual	actual/actual	No
Ireland	Annual	actual/actual	No
Italy	Annual	actual/actual	No
New Zealand	Semi-annual	actual/actual	Yes
Norway	Annual	actual/365	Yes
Spain	Annual	actual/actual	No
Sweden	Annual	30E/360	Yes
Switzerland	Annual	30E/360	No
United Kingdom	Semi-annual	actual/actual	Yes
United States	Semi-annual	actual/actual	No

Source: Choudbry (2004)

When determining the number of days in between two dates, include the first date but not the second; thus, under the actual/365 convention, there are 37 days between 4 August and 10 September. The last two conventions assume 30 days in each month, so for example there are 30 days between 10 February and 10 March. Under the 30/360 convention, if the first date falls on the 31st, it is changed to the 30th of the month, and if the second date falls on the 31st and the first date is on the 30th or 31st, the second date is changed to the 30th. The difference under the 30E/360 method is that if the second date falls on the 31st of the month it is automatically changed to the 30th. The day-count basis, together with the coupon frequency, of selected major government bond markets around the world is given in Table I.C.3.6.

I.C.3.5 Credit Risk

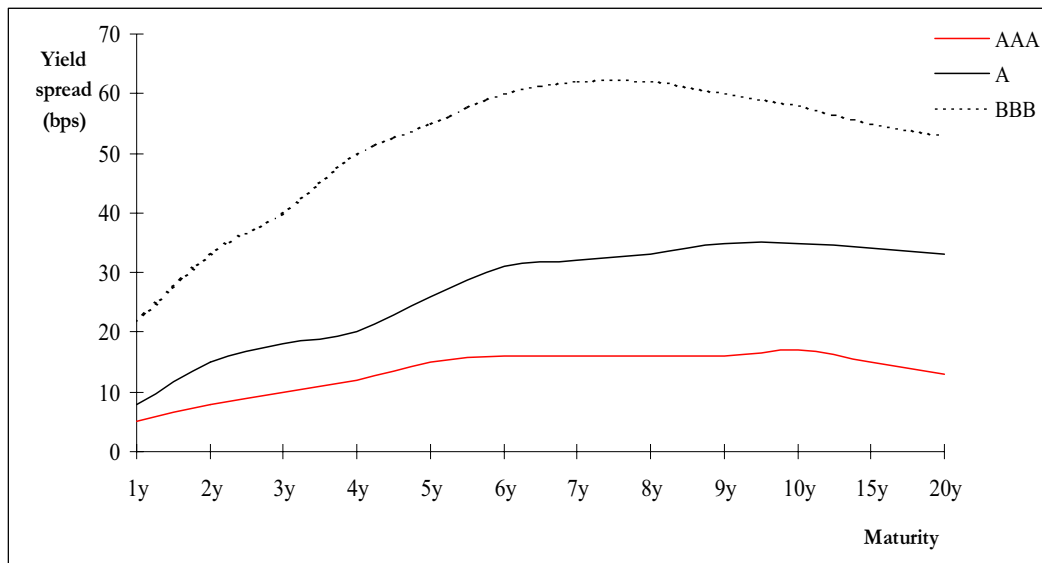
As is the case for government and municipal bonds, the issuer of a corporate bond has the obligation to honour his commitments to the bondholder. A failure to pay back interests or principal according to the terms of the agreement constitutes what is known as *default*. Basically, there are two sources of default. First, the shareholders of a corporation can decide to break the debt contract. This comes from their limited liability status: they are liable for the corporation's losses only up to their investment in it. They do not have to pay back their creditors when it affects their personal wealth. Second, creditors can prompt bankruptcy when specific debt protective clauses, known as covenants, are infringed.

In case of default, there are typically three eventualities:

- First, default can lead to immediate bankruptcy. Depending on the seniority and face value of their debt securities, creditors are fully, partially or not paid back thanks to the sale of the firm's assets. The percentage of the interests and principal they receive, according to seniority, is called the *recovery rate*.
- Second, default can result in a reorganisation of the firm within a formal legal framework. For example, under Chapter 11 of the American law, corporations that are in default are granted a deadline so as to overcome their financial difficulties. This depends on the country's legislation.
- Third, default can lead to an informal negotiation between shareholders and creditors. This results in an exchange offer through which shareholders propose to creditors the exchange of their old debt securities for a package of cash and newly issued securities.

A corporate debt issue is priced over the same currency government bond yield curve. A liquid benchmark yield curve therefore is required to facilitate pricing. The extent of a corporate bond's yield spread over the government yield curve is a function of the market's view of the credit risk of the issuer (for which formal credit ratings are usually used) and the perception of the liquidity of the issue. The pricing of corporate bonds is sometimes expressed as a spread over the equivalent maturity government bond, rather than as an explicit stated yield, or sometimes as a spread over another market reference index such as LIBOR. Figure I.C.3.3 illustrates some typical yield spreads for different ratings and maturities of corporate bonds.

Figure I.C.3.2: Yield spread by rating and maturity



Corporate bonds are much affected by credit risk. Their yields normally contain a default premium over government bonds, accounting for total default or credit risk, as well as over swaps. Swap spread, that is, the difference between the swap yield and the government yield with same maturity, is regarded as a systematic credit premium. In the four main bond markets swap yields reflect bank risk with rating AA, that is, the first rating grade below AAA, the normal rating for government bonds, accounting for specific default or credit risk.

Formal credit ratings are important in the corporate markets. Investors usually use both a domestic rating agency in conjunction with the established international agency such as Moody's or Standard & Poor's. As formal ratings are viewed as important by investors, it is in the interest of issuing companies to seek a rating from an established agency, especially if it is seeking to issue foreign currency and/or place its debt across national boundaries. Generally Eurobond issuers are investment-grade rated, and only a very small number are not rated at all.

Treasury securities are considered to have no credit risk. The interest rates they bear are the key interest rates in the US as well as in international capital markets. Agency securities' debt is high-quality debt. As a matter of fact, all rated agency senior debt issues are triple-A rated by Moody's and Standard & Poor's. This rating most often reflects healthy financial fundamentals and sound management, but also and above all the agencies' relationship to the US government. Among the numerous legal characteristics of the government agencies' debt, one can find that:

- agencies' directors are appointed by the President of the United States;
- issuance is only upon approval by the US Treasury;
- securities are issuable and payable through the Federal Reserve System;

- securities are eligible collateral for Federal Reserve Bank advances and discounts;
- securities are eligible for open market purchases.

Municipal debt issues, when rated, carry ratings ranging from triple-A, for the best ones, to C or D, for the worst ones. Four basic criteria are used by rating agencies to assess municipal bond ratings:

- the issuer's debt structure;
- the issuer's ability and political discipline for maintaining sound budgetary operations;
- the local tax and intergovernmental revenue sources of the issuer;
- the issuer's overall socio-economic environment.

I.C.3.6 Summary

In this chapter we have described the structure of the bond markets and the nature of the main market participants. These are comprised of issuers, investors and intermediaries.

Issuers are those who have a requirement to raise capital, and can be sovereign governments, government agencies, multilateral agencies and corporates. Bonds are rated in terms of their credit quality, which reflects the quality of the issuer. Certain sovereign issuers such as the USA, UK, Germany and Japan are viewed as triple-A or risk-free credit quality. Credit quality ranges from triple-A to D, meaning the issuer is in default.

Investors are fund managers, insurance companies, corporates and agencies. Bond markets are essentially over-the-counter markets, although a large number of them are listed on stock exchanges, such as the NYSE, Irish Stock Exchange and Luxembourg Stock Exchange. This listing enables institutional investors to hold them, who might otherwise be prevented from holding unlisted products.

Intermediaries include commercial and investment banks, who act as underwriters to bond issuances, as well as brokers and advisors. Note that banks also act as issuers and investors.

References

- Choudhry, M (2004) *Fixed Income Markets: Instruments, Applications, Mathematics*. Singapore: Wiley.
- Martellini, L, Priaulet, S, and Priaulet, P (2003) *Fixed-Income Securities: Valuation, Risk Management and Portfolio Strategies*. Chichester: Wiley.

I.C.4 The Foreign Exchange Market

Canadian Securities Institute¹

I.C.4.1 Introduction

The foreign exchange (or forex) market encompasses all the places in which one nation's currency is exchanged for another at a specific exchange rate. An *exchange rate* is the price of one currency in terms of another. For example, a Canadian dollar exchange rate of US\$0.63 means that it costs 63 US cents to buy one Canadian dollar.

The broadest definition of the forex market includes foreign currency purchases by individuals at bank branches for vacations or other personal reasons, as well as the large volumes exchanged between businesses and corporations and their respective banks. The largest component of the forex market, however, is the *interbank* market. The interbank market is an over-the-counter (OTC) market dominated by large financial institutions that buy and sell currencies among themselves. Some studies have estimated that between 90% and 95% of a bank's foreign exchange activity is with other participants in the interbank market.

In Section I.C.4.2 we examine the interbank market in more detail. Then, in Section I.C.4.3, we describe different types of exchange-rate quotations. We also explain how participants in the interbank market get quotes from one another, the role of the US dollar and how currencies are quoted relative to the US dollar, and how cross rates are calculated. In Section I.C.4.4 we provide a discussion of the different factors that affect foreign exchange rates. In Section I.C.4.5 we outline the differences between the spot and forward markets. An overview of a typical foreign exchange operation is provided in I.C.4.6 while I.C.4.7 concludes.

I.C.4.2 The Interbank Market

The interbank foreign exchange market has been described as a 'decentralised, continuous, open bid, double-auction' market. Let us look at these terms one by one.

Decentralised. The interbank market is an OTC market without a single location. It operates globally, through telephone and computer systems that link banks and other currency traders. Most trading activity, however, occurs in three major financial centres: London, New York and Tokyo. Smaller but important interbank centres exist in Frankfurt, Paris, Singapore and Toronto.

¹ Canadian Securities Institute, Toronto.

This decentralisation makes it extremely difficult to regulate the market. There is no true ‘regulator’ of foreign exchange markets, although the Bank for International Settlements (BIS) collects data on foreign exchange market activity. Forex markets are self-regulating, with associations such as the International Swaps and Derivatives Association playing an important role in fostering high standards of commercial conduct and promoting sound risk management practices. In addition, the capital adequacy requirements which apply to all financial institutions help to ensure that the risk of bank failure is minimised.

Continuous. Price quotes in the interbank foreign exchange market vary continuously. Banks call each other and ask for the current market price for a particular currency trade. They are quoted a bid/offer price that is stated as the price of one currency in terms of the other. The bid/offer price can change from moment to moment, reflecting changes in market sentiment as well as demand and supply conditions.

Open bid. Those who request a bid/offer price do not have to specify the amount they wish to exchange, or even whether they intend to buy or sell the currency. This is what is meant by an open bid: the bank that provides the quotation is open to buy or sell. The amount of the transaction is also left open, although conventional limits exist on what can be exchanged at the quoted rates. Typical price quotations are for trades worth US\$5–10 million or equivalent.

Double auction. Banks that receive calls for quotations also call other banks and ask for their market, that is, their buy (bid) and sell (ask) rates. The obligation to be both a price ‘maker’ and a price ‘taker’ is what is meant by a ‘double auction’.

In 2001, the BIS’s Triennial Foreign Exchange Survey showed that the global average daily turnover in the interbank foreign exchange market was US\$1210 billion. This is actually a decrease from the BIS’s previous survey in 1998, when daily turnover was nearly US\$1500 billion. The BIS cites four factors for the overall decline in activity:

- the introduction of the euro;
- consolidation in the international banking industry;
- consolidation in the corporate sector;
- a growing share of electronic brokering in the spot market.

I.C.4.3 Exchange-Rate Quotations

An exchange rate can be expressed relative to either of the two currencies involved. A *direct terms* quote is the number of *domestic* currency units that can be purchased with one foreign currency unit. An exchange rate of C\$1.5873 per US dollar is a direct terms quote from the perspective of a Canadian. From an American perspective, however, it is an indirect terms quote, meaning the number of *foreign* currency units that can be purchased with one domestic currency unit. An indirect terms quote expresses how much the domestic currency is worth in terms of the foreign currency – it is simply the opposite or ‘reciprocal’ of a direct terms quote. That is,

$$P_{B/A} = (P_{A/B})^{-1} = 1/P_{A/B}$$

where $P_{A/B}$ is the price of currency B in terms of currency A and $P_{B/A}$ is the price of currency A in terms of currency B.

Interbank participants may deal directly with one another or indirectly through a system of foreign exchange brokers or electronic brokering systems.

I.C.4.3.1 Direct Dealing

Direct dealing is the common practice by which a trader at one bank telephones a trader at another bank to get a quote on a certain currency. One of the advantages of this type of dealing is that a professional business relationship between the two traders develops over time, which could prove valuable in certain market situations. For example, if a certain currency is experiencing a temporary or even prolonged bout of illiquidity, an established relationship with another bank that deals in that currency may help a trader complete a deal at a reasonable price.

Banks control the amount of trading that they do with each other by placing limits on the amount of business that they will do with any other bank. The setting of these limits is a credit decision, made by senior risk managers. Banks limit their exposure to any given counterparty in order to minimise default risk, which is the risk that the other party to a trade cannot meet its contractual obligations on any given trade.

I.C.4.3.2 Foreign Exchange Brokers

Given the large number of banks participating in the global interbank market, or even within North America, Europe or Asia, one can imagine the intricate network of contacts and technology required to maintain relationships with as many other banks as possible. The use of brokers allows a bank to economise on its contacts with other market participants in one location. A foreign exchange broker acts as a middleman, bringing together two banks that have expressed an interest in buying or selling a currency at a specific price. For this service, brokers charge both the buyer and the seller a commission based on the size of the deal. Commission

rates are generally very similar across the board, but in certain cases discounts may be provided to banks that do very large trades.

Most brokers work in the large financial trading centres of London, New York and Tokyo, although some work in the smaller trading centres. Banks submit bids and offers for pairs of currencies to the brokers and the brokers are expected to disseminate or ‘work’ the orders to their respective networks. They do this by broadcasting the best bids and offers in various currencies over speakers that are physically located at a bank’s trading desk. When a trader wants to execute a deal based on the prices that are being broadcast, they will shout back to the broker via an open, direct telephone line, ‘mine’ (that is, I want to buy the currency) or ‘yours’ (I want to sell the currency).

Brokers will not reveal which banks are behind the prices until a deal is actually approved by both parties, assuming that both banks are able to deal with the other based on the limits they have imposed on their dealings. Brokers try to avoid matching banks that do not or cannot deal with each other. This is known as being aware of ‘who will take which names’ and which banks are ‘full’ on a given name.

The use of a broker guarantees anonymity to the buyer and seller. This is important to the trading process, because traders usually prefer not to reveal their position or market view to others in the market. Once a deal is concluded and both parties agree on the details, the deal is processed and confirmed by the respective back offices of the parties involved, including the broker. Brokers try to remain neutral in their dealings with banks and are not allowed to take positions the way banks can.

I.C.4.3.3 Electronic Brokering Systems

Until recently, ‘physical’ brokering – the process just described – was a key source of quotes and counterparties. However, due to the rapid evolution of technology and computing power, the physical brokering business is rapidly being supplanted by electronic brokering systems. Electronic brokering is similar to physical brokering, except that orders are placed into a computer system rather than with a person. The system automatically matches bids and offers as the price of a currency fluctuates and ‘market’ orders are matched with open orders.

The major providers of electronic brokering systems are EBS and Reuters. Other technology innovators have a smaller presence but are constantly trying to expand their share of the market as electronic brokering increases in importance.

I.C.4.3.4 The Role of the US Dollar

The US dollar has been the world's primary vehicle currency for almost a century. This means that it is the accepted benchmark currency against which most other currencies are valued, because most global trade transactions take place in US dollars. In addition, it is readily accepted as legal tender in some countries, and can be easily exchanged into the domestic currency in most others. Most currency trading in the interbank market therefore involves the US dollar on one side of the transaction – the BIS 2001 Survey showed that the US dollar is involved in 90% of interbank transactions.

Most currencies are quoted in indirect terms from the US perspective. For example, a 120.92 quote for JPY means that one US dollar is worth ¥120.92, or that it takes ¥120.92 to buy one US dollar. A JPY quote of 120.92 translates into a value of US\$0.00827 for one Japanese yen. If a currency is quoted in this manner, a rising quote signifies a strengthening of the US dollar relative to the other currency or a weakening of the currency relative to the US dollar.

Some currencies, however, are quoted directly from the US perspective, including the British pound sterling (GBP), the euro (EUR), the Australian dollar (AUD) and the New Zealand dollar (NZD). This type of quote indicates the value of the currency in terms of US dollars. For example, a 1.6234 quote for GBP means that one British pound is worth US\$1.6234, or that it takes US\$1.6234 to buy one British pound. A GBP quote of 1.6234 translates into a value of £0.6160 for one US dollar. If a currency is quoted in this way, a rising quote signifies strength in the other currency relative to the US dollar (and weakness in the US dollar versus the currency).

I.C.4.3.5 Market and Quoting Conventions

Traders have special ways of quoting bid and offer prices for foreign currencies. For example, a trader at bank A may call a trader at bank B and ask for B's market on the Canadian dollar (CAD) versus USD. B shows a market price of 1.5599–1.5604. Normally the trader at B would not waste his and the caller's time by saying 'one fifty-five ninety-nine' and 'one fifty-six-oh-four', but would rather quote only the last two decimal places, or points, as in 'ninety-nine' and 'oh-four'. Traders are always aware of what this means and will always know what the initial numbers are (often referred to as the 'big figure').

This quote means that the trader at B is willing to buy USD at C\$1.5599 and is willing to sell USD at C\$1.5604. In other words, A can either sell USD to B and receive C\$1.5599 or buy USD from B and pay C\$1.5604. These quotes are usually valid for amounts between US\$5 million and US\$10 million. The exact amounts are usually set up in advance. If a bank needs to trade a larger amount of money, the trader at bank A will specify this before the trader at bank B provides the

quote. If A buys USD at 1.5604 from B then it is paying or taking the offer. The bank A trader may simply say ‘mine’, which is interbank shorthand for ‘the US dollars are mine at your offer price’. If the trader at A wants to sell US dollars to B he or she will say ‘yours’, meaning the US dollars are yours at your bid price.

Table I.C.4.1 provides a typical list of quotes on several currencies relative to the USD. The bid–offer spread for most large, relatively liquid foreign currencies is usually 3–10 ‘points’. In times of extreme illiquidity, however, the spread can be significantly larger.

Table I.C.4.1: Typical foreign exchange quotes

Currency	Bid	Offer
CAD	1.5599	1.5604
GBP	1.5370	1.5375
EUR	0.9835	0.9839
JPY	117.50	117.60
CHF	1.4909	1.4915
AUD	0.5440	0.5449

Sometimes a trader may quote a ‘choice’ market, meaning the bid/offer price is the same and the interested party has the choice of buying or selling at this price. A choice market suggests extremely high liquidity in the market.

I.C.4.3.6 Cross Trades and Cross Rates

Most foreign currencies are not traded or quoted directly against one another. For example, if a corporation wants to sell Mexican pesos (MXN) in exchange for Hong Kong dollars (HKD), the transaction would take place as a cross trade that involves selling MXN for USD and then selling USD for HKD. The quote that is supplied for this trade would be derived from the two currencies’ quotes versus the US dollar. This quote is known as a cross rate.

Example I.C.4.1: Cross Rates

Suppose a bank is willing to buy and sell foreign currencies to and from its largest corporate customers at the exchange rates shown in Table I.C.4.1. If one of these customers wants a quote to buy CAD for JPY, the bank will arrive at a bid–offer spread for its cross rate as follows. The customer will sell CAD for USD at 1.5604. In other words, one Canadian dollar yields US\$0.6409. The customer will buy JPY for USD at 117.50. In other words, one Japanese yen costs US\$0.00851. Overall, one Canadian dollar will buy $0.6409/0.00851 = ¥75.30$.

If a customer wants to sell JPY for CAD, the cross rate would be calculated as follows. The customer will sell JPY for USD at 117.60. In other words, one Japanese yen yields US\$0.0085. The customer will buy CAD for USD at 1.5599. In other words, one Canadian dollar costs US\$0.6411. Overall, one Canadian dollar will cost $0.6411/0.0085 = ¥75.40$.

Altogether, the bank's bid–offer spread for the CAD in terms of JPY is 75.30–75.40. In other words, the bank is willing to buy Canadian dollars at ¥75.30 or sell Canadian dollars for ¥75.40.

I.C.4.4 Determinants of Foreign Exchange Rates

This section will give only a brief overview of the determinants of the value of a currency. Whole books have been written on how and why currency values change. Ask a forex trading professional why a currency is appreciating and sometimes, half-jokingly, he or she will say that there is more demand than supply. In fast-changing markets, when it is not immediately evident what is influencing currency value changes, this short answer is true, but it is not the whole story.

I.C.4.4.1 The Fundamental Approach

Foreign exchange rates have strong, long-term relationships with a country's identifiable economic fundamentals. These include: gross domestic product (GDP); rate of inflation; productivity; interest rates; employment levels; balance of payments; and current account balance. The performance of a national economy is generally measured by changes in its GDP. From quarter to quarter, and from year to year, various statistical agencies report on this measure. If the economy of one country is performing well relative to that of another country, the country with the stronger economic performance will usually have a stronger currency. Comparisons of the economic strength of countries are most often conducted relative to the United States.

Currencies are affected not only by what has happened in the past but also by expectations and forecasts of future performance. Attempts to anticipate the future movement of a currency are also usually done relative to the US dollar, because most currencies trade directly against the dollar. It is difficult to anticipate where the Swedish krona will be relative to the Mexican peso, but both these currencies can be compared to the American dollar. From there it is possible to come up with a calculated or projected cross rate, but this rate will have a larger margin for error.

Real GDP is related to inflation levels. Higher levels of inflation will erode the domestic value of the currency, because it now takes more domestic currency to buy foreign goods. Higher levels of inflation also reduce the real level of interest rates. Foreign investors will have less incentive to invest in a country with high inflation.

The productivity of a country also affects the value of the currency. For instance, when productivity increases because of technological advances, this improvement translates into stable prices and low inflation. The currency market rewards countries with high levels of productivity, as it did the United States in the mid- to late 1990s. Many economic analysts believe that this was a key reason for the strong US dollar during the 1990s.

Real interest rates also help determine the value of a currency. For instance, if the Bank of Canada advocates a 'tight' monetary policy by raising interest rates, then (all things being equal) this policy will attract more foreign investment in Canadian money assets. Conversely, if monetary policy is 'loose', foreign investors may abandon the Canadian dollar in favour of higher-yielding currencies.

The level of employment in a nation also influences the value of a currency. When employment levels are high, consumer and general household consumption will be strong and the economy will benefit. During much of the 1990s, the US economy performed relatively well. Good employment opportunities and consumer purchases helped prevent the US economy from slowing down and bolstered the value of the US dollar.

A country's balance of payments helps determine the value of a currency. The larger the balance of payments, the stronger the currency value relative to other world currencies. A major component of the balance of payments equation is the current account balance. While the current account balance is derived from the value of a country's merchandise and service imports and exports, as well as the flow of interest and dividends to and from the country, it can also be related to the level of savings and investment in the country and the government's budget balance. In general, the greater the government's surplus, the greater the current account balance. When the current account balance is positive and the trend is positive, the domestic currency will be strong. It makes sense that as governments move from deficits to surpluses, the value of the currency will rise.

The above list is by no means exhaustive and other factors can influence the value of a currency. Globalisation has increased the number of influences on the currency of each country. Forecasters must consider the effects of dozens of factors that may influence the value of a currency. This can be a difficult and tedious process with no guarantee of success or accuracy.

I.C.4.4.2 A Short-Term Approach

Many interested parties participate in the foreign exchange market, including both passive and active stakeholders. Foreign exchange professionals, including traders and account executives at banks, all watch the above factors to discern long-term trends in the value of a currency. However, foreign exchange traders tend to have a shorter-term view and work within certain limits. They have to meet profit targets that may be daily, weekly or monthly, so they react to events that influence a currency in the shorter term.

The forex market is fast-moving and volatile. On any given day, a currency may move a few points or several hundred points. Traders try to take advantage of these changes, large or small, to make a profit. Traders consult charts and use technical analysis to understand currencies the same way that equity analysts and traders try to understand the position and future of a particular company or industry. Charts show short-term patterns that may help a trader make a decision.

One factor that influences the value of a currency in the short term is the general level of liquidity. Markets usually exhibit a certain degree of fluidity, but at times liquidity dries up. During these times, currency values may be quite volatile. Speculators enter the market and try to push the currency in a certain direction to suit their position. Movements can be abrupt and exaggerated, and traders must take care not to get trapped in this environment with an unfavourable position. Fortunately, this situation is usually temporary.

The interplay between the futures and the cash markets can affect currency values. Usually, the cash market leads the futures market. However, at times this pattern is reversed for a very brief period. This reversal creates short-term disequilibrium and markets become very volatile until liquidity is restored. Temporary opportunities may open up, creating a trading frenzy for a short period.

Traders often try to manipulate or push the market to achieve a more favourable position. For example, a large institution, usually a bank, may hold a certain position in the foreign exchange OTC derivatives market. Suppose a bank is short a CAD/USD call option with a strike price of 1.5000 and a knock-out feature at 1.5800 – meaning that the call option becomes void if during the life of the contract the spot price hits 1.5800. If the current spot market is at 1.5650, the bank's trader may try to push the market above 1.5800 to avoid the potential payout to the buyer of the option.

Unexpected political events or world crises can affect currency values. In a crisis, investors look for safe places to invest and certain currencies, such as the US dollar or the Swiss franc, are

considered safe havens. Money will flow from weaker second-tier currencies to these safe havens. These markets are temporary and impossible to predict, but can have devastating long-term results for an individual market participant. In fast-changing markets, only the bravest or most foolhardy participants come out to play; most business is done on an as-needed basis.

I.C.4.4.3 Central Bank Intervention

Central banks (CBs) carry enormous clout in the foreign exchange markets but they usually exercise their power cautiously. However, from time to time, CBs intervene in the foreign exchange markets. Some CBs are more active than others and levels of transparency vary. Most CBs try to keep a variety of options available to ensure the greatest impact when they do intervene.

CBs generally let the market determine the level of a currency. Most CBs in industrialised countries work to foster sustainable economic prospects and keep prices stable, rather than dictating or managing the level of the home currency. From time to time, however, CBs do intervene in the foreign exchange market. The intervention can take the form of what is known as *moral suasion*, whereby a CB spokesperson may offer the markets the CB's views on the current value of the home currency. These views reveal the CB's preferred position for the currency. Usually, the market acts to adjust the currency value to the desired levels to prevent direct intervention from the CB. However, market participants occasionally take positions opposite to the CB's preference. This can be a dangerous game, as the consequence may be high for both the CB and these contrary market participants. The CB may lose face because its credibility is at stake. The contrary market participants risk big losses if the CB decides to intervene.

CBs intervene from time to time when markets get disorderly. Opinions may differ as to whether a market is behaving in a disorderly way, but CBs like the Bank of Canada employ professional forex personnel who can interpret market conditions so that intervention is accurate and effective and stability can be restored. Some interventions are focused on keeping a certain pair of currencies within a certain range. For instance, the Bank of Japan tends to state openly where it would like to see the value of the Japanese yen relative to the US dollar. Market participants know the Bank's biases and try to avoid provoking the Bank of Japan for fear that it will intervene directly. CBs also communicate with their counterparts in other countries to monitor conditions. They may ask their counterparts to intervene on their behalf if the situation is sufficiently serious.

In summary, CBs in industrialised nations intervene in the marketplace to influence the short-term movement of currencies. This intervention is different from the post World War II to 1971

Bretton Woods regime, under which intervention by a CB was an official proclamation of a structural change in the value of a currency. Today, the general philosophy is that the markets will ultimately dictate currency equilibrium.

I.C.4.5 Spot and Forward Markets

A currency can be bought or sold in either the spot market or the forward market.

I.C.4.5.1 The Spot Market

The interbank spot market consists of purchases and sales of currency for *immediate delivery*. For CAD and USD transactions, ‘immediate delivery’ occurs on the next business day after the trade is completed. For most other currencies, including the major European and Asian currencies, ‘immediate delivery’ occurs two business days after the trade date. Each currency transaction involves two sets of payments, one for each of the two currencies involved in the trade.

Suppose that on Monday a trader at BMO buys US\$10 million spot from a trader at CIBC at a CAD exchange rate of 1.5604. To settle this transaction, CIBC will notify its correspondent bank in New York to debit its US dollar account by US\$10 million and send the money to BMO’s correspondent bank in New York for further credit to BMO.² The transfer of US dollars through the correspondent banks is done through the Clearing House Interbank Payments System (CHIPS), a central clearing house for USD transactions conducted by its member banks. The CHIPS transfer will settle the next business day, on Tuesday (assuming both Monday and Tuesday are US business days).

At the same time, BMO will send CIBC C\$15,604,000 through the Large Value Transfer System (LVTS) operated by the Canadian Payments Association (CPA). BMO will use the Society for Worldwide Interbank Financial Telecommunications (SWIFT) system to transmit the instructions. At the end of the day, the LVTS balance for each CPA member is settled by a debit or credit to its account with the Bank of Canada. If this forex trade is the only transaction between CIBC and BMO on this day (an unrealistic assumption), then BMO’s account with the Bank of Canada will be debited by C\$15,604,000 and CIBC’s would be credited for the same amount, all for settlement on Tuesday.

² A correspondent bank is a member of a national payments clearing system that clears and settles transactions on behalf of a foreign customer. Each bank involved in the interbank market has at least one correspondent bank in each country in which it conducts forex transactions. All of a bank’s foreign exchange balances are held by its correspondent banks on its behalf.

I.C.4.5.2 The Forward Market

The main thing that distinguishes the forward market from the spot market is the timing of delivery. Spot market transactions settle one or two business days after the trade date, but the settlement of forward market transactions can occur from one week after the trade date to as much as 10 years after the trade date. Because of this delayed settlement, forward prices are different from spot prices. In other words, not only is the time of delivery between a spot and forward transaction different, but usually the price is too. The principles of currency forward pricing are discussed in Section I.B.3.3

Liquid forward markets exist in the major currencies for one-month, two-month, three-month, six-month and one-year delivery dates.³ Longer delivery periods are possible for certain pairs of currencies (such as CAD/USD). All the pertinent details of the trade – the price (exchange rate), the size of the trade, and the settlement procedures – are agreed to at the time of the trade. This commitment to trade currencies at a previously agreed exchange rate is known as a forward contract.

The characteristics of trading in the spot interbank market also apply to trading in the forward interbank market. That is, the market is a decentralised, continuous, open-bid, double-auction market. Most currencies are quoted relative to the US dollar in either American or European terms. Transactions are conducted either directly between bank traders, or through physical or electronic brokering systems. Forward transactions are settled just like spot transactions, and involve a transfer of currencies in two different countries.

I.C.4.5.2.1 Forward Discounts and Premiums

Apart from the delivery date, the other major difference between spot and forward transactions is the price at which forward trades occur. The absolute difference between the spot and forward price of a currency is called the *currency swap rate* or, simply, ‘swap points’. The relative annualised difference is known as a *forward premium* or *forward discount*, depending on whether the forward price F_n is higher or lower than the spot price S :

$$\text{Swap rate or Swap points} = F_n - S, \quad (\text{I.C.4.1})$$

$$\text{Forward premium or Forward discount} = \frac{F_n - S}{S} \times \frac{360}{n}. \quad (\text{I.C.4.2})$$

³ The actual delivery date is the number of months after the spot delivery date. For example, a one-month CAD/USD forward settles in one month and one day.

Example I.C.4.2: Currency Swap Rates and Forward Premiums

Suppose the CAD/USD spot price is 1.5870 and the one-month (30-day) CAD/USD forward price is 1.5884. Then the one-month CAD/USD swap rate is $1.5884 - 1.5870 = 0.0014$. Traders would call this a swap rate of 14 points.

The annualised one-month forward premium (or forward discount, if the forward price is less than the spot price) is:⁴

$$\frac{1.5884 - 1.5870}{1.5870} \times \frac{360}{30} = 0.010586 = 1.06\%.$$

Since the CAD/USD quote is the price of the US dollar in terms of Canadian dollars, we say the US dollar is trading at a one-month forward *premium* of 1.06% relative to the Canadian dollar. Conversely, we could say that the Canadian dollar must be trading at a one-month forward *discount* of approximately the same amount – ‘approximately’, because the process of inverting currency quotes alters the exact value of the premium or discount. If the inversion is calculated to several decimal places, the difference will be small. If it is rounded to three or four decimal places, the difference will be larger.

For example, if we invert the spot and forward CAD/USD quotes above and round them to four decimal places, we get a spot value of 0.6301 and a forward value of 0.6296. Plugging these values into (I.C.4.2) gives

$$\frac{0.6296 - 0.6301}{0.6301} \times \frac{360}{30} = -0.00952 = -0.95\%,$$

that is, a discount of 0.95%. Here the difference (1.06% compared to -0.95%) seems quite pronounced. But if we repeat this calculation using, say, eight decimal places, 0.63011972 and 0.62956434, the difference is negligible:

$$\frac{0.62956434 - 0.63011972}{0.63011972} \times \frac{360}{30} = -0.010577 = -1.06\%.$$

I.C.4.5.2.2 Interest-Rate Parity

What determines the forward exchange rate – and hence the swap rate and the forward discount or premium? As with most other financial forward contracts, the forward price is derived from the spot price, based on the cost-of-carry model. If it is not, arbitrage can produce risk-free profits. The arbitrage transactions that keep forward exchange rates in line with spot exchange

⁴ See Chapter I.C.3 for the correct day-count conventions for interest rates. Here we use 360/30 for the CAD/USD quotes.

rates are known as covered interest arbitrage. In foreign exchange trading, the effect of covered interest arbitrage is known as interest-rate parity.

Interest-rate parity means that the currency of a country with a low interest rate should trade at a forward premium relative to the currency of a country with a high interest rate. Interest-rate parity effectively eliminates the interest-rate differential between countries after foreign exchange risk has been eliminated with a forward contract.

When interest-rate parity holds, the *covered interest differential* – that is, the difference between the interest rate in one country and the interest rate in another country, combined with a forward contract – is zero. Put another way, the interest-rate differential should be approximately equal to the forward discount or premium. If it is not, covered interest arbitrage by interbank and large institutional traders will quickly eliminate the interest-rate differential. Interbank and large institutional traders focus on Eurocurrency interest rates, because they can easily borrow and lend large amounts of money in this market.

Example I.C.4.3: Covered Interest Arbitrage

Suppose that the spot CAD/USD exchange rate is 1.5870 and that three-month (91-day) interest rates in the Eurocurrency market are 3.5% for Canadian dollars and 3% for US dollars. Also, suppose that the three-month CAD/USD forward rate is 1.5870, the same as the spot price. Since the Eurocurrency rates for US dollars are lower than they are for Canadian dollars, the US dollar should be trading at a forward premium relative to the Canadian dollar. Since it is not, covered interest arbitrage will produce a risk-free profit. There are four steps in the covered interest arbitrage:

1. Borrow US dollars. If US\$1 million is borrowed in the Eurodollar market at 3%, the amount that must be repaid at the end of three months is:

$$\text{\$1 million} + \left(\text{\$1 million} \times 0.03 \times \frac{91}{360} \right) = \text{\$1,007,583.}$$

2. Sell US dollars and buy Canadian dollars spot. Convert the US\$1 million to Canadian dollars at the spot rate of 1.5870 for proceeds of C\$1,587,000.

3. Invest the Canadian dollars. Invest the Canadian dollars in the Eurocurrency market for three months to earn 3.5%. After three months, the Canadian dollar investment will be worth:

$$\text{C\$1,587,000} + \left(\text{C\$1,587,000} \times 0.035 \times \frac{91}{360} \right) = \text{\$1,601,040.}$$

4. Sell Canadian dollars and buy US dollars forward. Sell C\$1,601,040 forward at the three-month CAD/USD forward rate of 1.5870 for proceeds in three months' time of $1,601,040/1.5870 = \text{US}\$1,008,847$.

In three months, the proceeds from the Canadian dollar investment can be used to satisfy the obligations of the forward contract and in return will produce US\$1,008,847, of which US\$1,007,583 must be used to repay the USD loan. After all the transactions have settled, US\$1264 is left as a risk-free profit.

In the example above, when other interbank and institutional traders realise that covered interest arbitrage offers a risk-free profit, they too will engage in these transactions. This will have the following effects: US dollar interest rates in the Eurocurrency market rise as traders borrow US dollars; the CAD/USD spot rate falls as traders sell US dollars spot; Canadian dollar interest rates fall as traders invest Canadian dollars; the CAD/USD forward rate rises as traders buy US dollars forward. Eventually, all four prices and rates will converge on values that reduce any chance of a risk-free profit when carrying out covered interest arbitrage.

The following equation can be used to determine the fair value forward exchange rate that eliminates risk-free profits from covered interest arbitrage, assuming equal borrowing and lending rates and no bid–ask spreads:

$$F_{X/Y} = S_{X/Y} \frac{1 + r_X n / 360}{1 + r_Y n / 360}, \quad (\text{I.C.4.3})$$

where $F_{X/Y}$ is the n -day forward exchange rate for currency Y in terms of currency X, $S_{X/Y}$ is the spot exchange for currency Y in terms of currency X, r_X and r_Y are the interest rates on currencies X and Y respectively, stated as money market yields in decimal form, and n is the number of days covered by the two investments and the forward contract.

Example I.C.4.4: Fair Value Forward Rates

Based on a spot CAD/USD exchange rate of 1.5870 and 91-day Eurocurrency rates of 3% for US dollars and 3.5% for Canadian dollars, the three-month forward rate should be:

$$1.5870 = S_{X/Y} \frac{1 + 0.035 \times 91 / 360}{1 + 0.03 \times 91 / 360} = 1.5870 \frac{1.008847222}{1.007583333} = 1.5890.$$

Returning to the covered interest arbitrage example above, if the forward rate had been 1.5890 rather than 1.5870, the proceeds from the forward contract in step 4 would have been $1,601,040/1.5890 = \text{US}\$1,007,577$. This is actually US\$6 less than we need to repay our US dollar loan, which would remove the incentive to engage in these transactions in the first place.

Remember that when interest-rate parity holds, the annualised forward discount or premium should be approximately equal to the interest-rate differential. Based on a forward price of 1.5890, the annualised forward premium on the US dollar is equal to:

$$\frac{1.5890 - 1.5870}{1.5870} \times \frac{360}{91} = 0.004986 = 0.499\%.$$

This is almost exactly equal to the 0.5% differential between US dollar and Canadian dollar interest rates. In other words, the premium paid to lock in a forward purchase of US dollars will effectively eliminate the higher interest rates earned by investing in Canadian dollars.

Example I.C.4.5: Hedging in the Forward Markets

A Canadian firm imports components for its manufacturing process from the United States. The company is expecting a large shipment three months from now, at which time a payment of US\$3,000,000 will be required. Concerned about a possible appreciation in the USD, the CFO decides to hedge this short USD position.

Using the rates from the previous example, the CFO can buy US dollars at 1.5890 for delivery in three months. Once the forward contract is in place, the CFO knows with certainty that the shipment of components will cost C\$4,767,000 ($3,000,000 \times 1.5890$), regardless of exchange-rate movements in the forthcoming months.

I.C.4.6 Structure of a Foreign Exchange Operation

Trading rooms around the world are set up in a similar fashion. Although the number of employees and floor space may vary, a large trading operation in London, Tokyo or New York will look very similar. Most foreign exchange operations have the following structure. A spot desk consists of several employees who trade various currencies on a spot basis in the interbank market. The domestic currency most likely is the focal point of the whole trading operation and a chief or senior trader is responsible for generating a certain profit figure for the trading operation. This proprietary trading uses the bank's capital and a certain return is expected on the capital employed. Depending on the risk profile, certain traders may be assigned to work with other currencies that represent the bank's interests or for which the bank has economies of scale. For example, many banks trade euros, yen, pounds sterling and Swiss francs because of the depth and liquidity of these markets. Some banks employ traders to work with second- or third-tier currencies. These 'exotics' are typically tied to customer transactions; although the trades are smaller, they can be very profitable.

The spot desk is complemented by a forward desk, staffed by employees who trade and manage the bank's forward currency positions. Depending on its scope of operation, the forward desk may also be involved in other banking operations, such as the money market or funding desk. A bank may try to fund its loan obligations by trading in currencies other than the domestic currency; this is where the forward desk plays a key role. The forward desk also manages the bank's cumulative forward position derived from the activities of the bank's corporate clients.

A chief trader usually manages the spot desk and the forward foreign exchange desk. These two desks are responsible for trading, as opposed to advisory services to clients. The compensation paid to trading professionals in this area is usually tied to the profits generated by their trades. A medium-sized or larger forex trading operation may also include a forex derivatives desk. This desk would be responsible for trading exchange-traded or OTC options and currency futures. Most banks have a section within the forex department known as *foreign exchange advisory services*, sometimes called a sales desk. The role of the advisory desk's account executives is to attract new forex business or clients to the bank, provide professional services to existing clients, and provide pricing for clients. Sales executives work closely with other officers of the bank, especially credit personnel, to ensure that deals with clients remain within prescribed credit limits. The sales desk can be broken down into three subcategories: institutional coverage to larger, more sophisticated accounts, such as governments or money managers; corporate or commercial accounts; and the retail business from the branch network.

All these desks together constitute the *front office*. A healthy conflict often exists between the trading and sales desks because of different objectives related to the pricing of forex products. Sales executives always want competitive rates for their existing clients or for prospecting opportunities, while the traders need to maximise profits.

All the dealing operations in the front office are supported by *back office* departments. The back office processes the transactions, including record keeping, applying checks and balances, ledger and sub-ledger activity accounting, reconciliations, deal confirmations and deal settlements. In smaller operations, the back office may also be required to monitor the dealing limits of the foreign exchange traders to ensure compliance.

The increasing sophistication of financial products and some high-profile financial fiascos have led to the evolution of an area often referred to as the *middle office*. This group is essentially a risk management group mandated to ensure compliance within the trading room. The risk management group identifies, quantifies, monitors, and analyses the risk–reward profile of a trading operation in terms of market, liquidity, credit and operations.

Front office (i.e. trading and sales) and back office objectives are not the same. The former is profit-motivated, while the latter is focused on checks and balances. The personnel in these two different groups usually report to different senior officers of a bank to ensure that no conflict of interest arises.

Foreign exchange operations vary from firm to firm, with differences mostly in degree as opposed to kind. For instance, a smaller forex group may have some crossover in responsibility, whereby a spot trader also trades forex derivatives or a trader of secondary currency also trades the forward book. Some organisations employ a 'corporate dealer' who carries out both trading and advisory functions.

It is important to distinguish the responsibilities of a forex trader from those of an account executive. The forex trader is paid to take positions in the marketplace, but the account executive is not. Although this distinction may vary from one firm to another, advisory personnel are generally not supposed to take positions either from the market directly or from clients. Although there is no specific academic requirement for the position of account executive, most account executives have some post-secondary education. In addition, management encourages account executives to enrol in industry courses to gain additional insight and knowledge of foreign exchange markets and other aspects of the business. Banks tend to be fairly generous in their financial support for employees. Compensation for foreign account executives and traders usually has a fixed component and a variable bonus, depending on how well the department performs over the budget year. Forex professionals in banks do not work on commission.

I.C.4.7 Summary/Conclusion

An exchange rate is the price of one currency in terms of another. The interbank foreign exchange market is known as a decentralised, continuous, open-bid, double-auction system because it operates through telephone and computer systems that link banks and other currency traders around the world; price quotes in the interbank foreign exchange market vary continuously during the trading day; the trader who requests a quote does not have to specify the amount he or she wishes to exchange, or even whether he or she intends to buy or sell the currency; and banks that receive calls for quotes also call other banks and ask for their market. A direct terms quote is the number of domestic currency units that can be bought with one foreign currency unit, while an indirect terms quote is the number of foreign currency units that can be bought with one domestic currency unit. Direct dealing describes the common practice by which a trader at one bank telephones a trader at another bank to get a quote on a certain currency. A foreign exchange broker acts as a middleman, bringing together two banks that have expressed an interest to buy or sell a currency at a specific price. In electronic brokering, orders are placed

into a computer system rather than with an individual broker. The system automatically matches bids and offers as the price of a currency fluctuates and as 'market' orders are entered.

Most currency trading in the interbank market involves the US dollar on one side of the transaction. If we take the US perspective, most currencies, including the Canadian dollar, are quoted indirectly against the US dollar. A few currencies, including the British pound and the euro, are quoted against the US dollar directly.

A cross rate is a quote for a foreign currency against another foreign currency, without reference to the US dollar. It is important to use the correct quote relative to US dollar when determining the bid and offer quote of a foreign currency relative to another foreign currency. If the party asking for the quote wants to buy currency A in exchange for currency B, the provider of the quote must use an offer quote on currency A and a bid quote on currency B. The actual quote may be stated in terms of either currency A or currency B.

Low levels of liquidity may exaggerate short-term movements in the currency in one direction or the other. The currency futures market can lead the cash market for brief periods of time. Traders can try to manipulate prices. Unexpected political events or world crises can affect currency values. A central bank can influence the value of its currency by moral suasion or by directly buying or selling the currency in the market.

Currency transactions involves two sets of payments, one for each of the two currencies involved in the trade. Banks in one nation settle their foreign currency transactions through correspondent banks in the foreign country. Foreign exchange transactions involving a bank's domestic currency will settle through that nation's domestic payments system. Spot market transactions settle one or two business days after the trade date, while the settlement of forward market transactions can occur from one week to 10 years after the trade date. This delayed settlement usually leads to forward prices that are different from spot prices.

The forward premium or forward discount between two currencies is the annualised difference between the spot and forward exchange rates. Interest-rate parity means that the currency of a country with a low interest rate should trade at a forward premium relative to the currency of a country with a higher interest rate. Interest-rate parity effectively eliminates the interest-rate differential between countries after foreign exchange risk has been eliminated with a forward contract. Covered interest arbitrage forces the forward discount or forward premium on a currency to approximately equal the interest-rate differential between the two currencies.

I.C.5 The Stock Market

Dr Andrew Street¹

I.C.5.1 Introduction

Stocks (also known as *shares* or *equities*) represent an interest in the ownership of companies or corporations. These securities may exist as paper certificates ('bearer form') or notional entries in the computers of the share register ('book entry form'). These stocks are bought and sold (traded) among different market participants, including investors, hedge funds and investment banks. Stock are issued and sold by companies or corporations on their formation as a way of raising working capital and spreading the risk of ownership among shareholders according to their individual risk appetite. Companies and corporations are often founded as small private ventures with a limited number of shareholders who know each other and are often directly involved in the business. As the business grows and the need for capital expands, the existing shareholders frequently 'float' the company on the stock market by issuing new shares to new investors. This typically dilutes their shareholding and they often sell some of their own stake at the same time. This process is often known as 'listing' or doing an initial primary offering (IPO). Alternatively, as corporations grow, they may need more capital. One way of raising extra funds for a company that is already listed is to make a 'secondary' or 'rights offering'. Here existing shareholders are offered the right to subscribe to new shares in a corporation, usually at a substantial discount to the current market price.

Access to the stock market is regulated in most developed economies so that a company applying for a listing has to achieve and maintain minimum standards of capitalisation, disclosure and financial standing. The other market participants, such as professional traders, investment banks and investors, are also required to operate within the rules of the market, which may include refraining from activities such as 'insider trading'.

A stock market is therefore, in general, a regulated marketplace for the buying and selling of the ownership of corporations for the purpose of spreading risk and raising capital. The corporations benefit by having a large liquid market in which to raise capital, and investors benefit by having the ability to spread and control their investment risk via the liquidity that a large and deep market offers. Intermediaries such as investment banks benefit by generating commissions and fees on stock trades and participation in the movement in price of the individual stocks.

¹ Managing Director, Value Consultants Limited, London.

The details of stock that are listed on the various stock markets are generally available via daily official lists from the market controller (e.g. the London Stock Exchange or New York Stock Exchange) which are often reproduced in whole or part in newspapers such as the *Financial Times* or *Wall Street Journal*. This information is also invariably available electronically via data vendors such as Reuters and Bloomberg and includes statutory disclosures (part of the listing requirement) by the listed company of information affecting shareholders, such as dealings in shares by a director of that company.

The market may itself have divisions into ‘senior’ or ‘junior’ listings with higher or lower listing requirements, and therefore potentially more or less investment risk. For example, the UK has the Alternative Investment Market, which has less onerous listing and capitalisation requirements than a full stock exchange listing. Other stock markets, such as Luxembourg or Johannesburg, may have higher or lower listing standards and requirements than, say, London or New York. Some very large companies may be listed on more than one stock exchange: for example, HSBC is listed in Hong Kong and London. The total value of stocks traded on the world’s stock markets is approximately US\$30 trillion dollars ($\text{US\$}3 \times 10^{13}$) with the USA having the largest single market (approximately 50% of the total), followed by the Eurozone, UK and Japan.

In the following sections we will look in more detail at the characteristics of the stock market and its participants, the properties of common equities, the primary and secondary markets, and the mechanics of trading, including costs, strategies (such as going short), and the use of leverage via margin trading or exchange-traded derivatives.

I.C.5.2 The Characteristics of Common Stock

As we have seen above, common stocks represent the ownership of a company or corporation. More specifically, the equity holder has a *claim on the residual assets* of a company after all other claims have been paid in the event of a liquidation of the company. If we consider a simplified company balance sheet, we can illustrate this point more easily (Table I.C.5.1).

Table I.C.5.1: Simplified balance sheet of a company

Assets	Liabilities
Fixed Assets	Creditors/Bond Holders
Debtors	Shareholders’ Funds

The balance sheet equates assets to liabilities; the balancing item is the shareholders’ funds. The shareholders’ funds represent the value that the equity shareholders would be entitled to if the company were liquidated. Clearly, the greater the assets of the company and the smaller the

creditors, the greater the value of the company and therefore the greater the value of the shares to the shareholder. However, practically speaking, this value is directly accessible only by breaking up or liquidating the company. When this happens, there is a strict hierarchy of pay-out (see Table I.C.5.2) to the creditors of the firm based on seniority.

Table I.C.5.2: Pay-out hierarchy on liquidation

<i>First in the queue</i>	Inland Revenue/Tax Authorities
	Secured Creditors (Mortgagor)
	Trade Creditors
	Senior Bond Holders
	Junior Bond Holders
<i>Last in the queue</i>	Equity Holders

This means that senior creditors such as the Inland Revenue or the bond holders will be paid in full prior to any remaining assets being distributed to the equity holders. In this sense the equity holders have a *residual claim* on the company based on seniority. They are the most *junior* and therefore the *bearers of the most risk* on liquidation. Generally, this liquidation or break-up of the company would occur only when the company became insolvent and could no longer operate. In those circumstances it is unlikely that the residual value available to shareholders would be significant and is likely to be close to zero. This explains why, when a company announces a serious operating problem or potential insolvency that is not widely known, the equity share price rapidly falls towards zero. Similarly, when the company has a sudden windfall success (e.g. an oil field find or successful drug trials) the equity price often jumps substantially.

I.C.5.2.1 Share Premium and Capital Accounts and Limited Liability

Shares usually have a nominal value, typically in the UK 25p. These shares are known as ‘ordinary shares of 25p nominal value’. At primary issue the shares are often sold at a premium, say £1, in which case 25p of the sale price goes into the *share capital account* of the shareholders’ funds and 75p goes into the *share premium account*. Both of these accounts, along with the profit-and-loss account, constitute the shareholders’ funds in the company balance sheet. Under certain accounting/regulatory circumstances, the ‘surplus’ in the share premium account can be used by the company to offset some costs such as ‘goodwill’ on the purchase of other companies. After the initial issue of shares, they then trade in the secondary market at the ‘market’ price. This price then simply reflects the amount agreed by both counterparts to buy a share and this amount is paid to the seller in exchange for the equity share.

Once the shares are sold in a primary issue, the equity holder's liability to the debts of the company is limited to the amount already paid for the shares, i.e. the value of the share premium and share capital accounts. Occasionally, shares are issued 'part-paid', which means that only a fraction of the issue price is paid upfront, with further payments due in the future, or, in the event of company default, due immediately. In this sense the owner of the company, the equity shareholder, has *limited liability* to the debts of the company, the amount of this liability being the fully paid-up share capital. This is the fundamental idea of a 'limited-liability corporation'.

I.C.5.2.2 Equity Shareholder's Rights and Dividends

The ordinary shareholder usually has voting rights associated with his holding in the corporation's stock at special company meetings such as the annual general meeting (AGM). Therefore, a shareholder with 30% of the issued equity would carry a 30% weight in any vote or resolution. This means that decisions regarding the company's management and overall direction are controlled by ballot weighted by ownership. There are exceptions where voting rights are not evenly distributed throughout the equity shareholders, although these are increasingly rare.

The shareholders are responsible for electing the company's board of directors, who in turn control the day-to-day activities of the company. The AGM provides a regular forum for shareholders to air their views and ultimately control their company by the support (or otherwise) of the managing director and the rest of the company board of directors. The directors run the company on behalf of, and with the permission of, the shareholders, who, with sufficient voting numbers, can remove them at any time.

In addition to voting rights, the equity shareholders will also receive dividend payments on their stock if the board of directors decide that a dividend can be paid. This decision is based on the financial standing of the company. Dividends are declared annually and may be varied or stopped completely if the directors decide that it is in the company's interest to do so. Typically, in the UK, dividends are paid twice a year (usually an *interim* and then a *final dividend*, the former usually being smaller); in the USA dividends on large companies are paid four times a year in equal instalments. Dividends are usually announced and a record and payment date set, such that all holders of the equity on the record day qualify for the dividend payment. After that date the share trades 'ex-dividend' and the payment of the dividend is made at some later date to the holder on the record date, who by then may have sold his or her shares on as 'ex-dividend' stock.

For instance, suppose Glaxo Ordinary 25p shares are trading at £12.50 mid (£12.45 bid, £12.55 offered) and a final dividend of 25p is announced. The record date is 15 September and pay day is 5 October. If an interim dividend of 15p has already been paid six months earlier, the total

dividend payment for the year is 40p. Then the simple annual dividend yield is $(40/1250) \times 100\% = 3.20\%$.

Therefore the holders of common stock have voting rights to control their company and they receive dividends if they are considered appropriate for the company by the board of directors. The board is ultimately controlled via the shareholders' votes. The return on an investment in equity is a combination of dividends received and any capital appreciation (increase in price) realised when the stocks are eventually sold. Equity shares can therefore provide a combination of income (which is uncertain) and capital gain (which is also uncertain). This makes equity valuation relatively difficult!

I.C.5.2.3 Other Types of Equity Shares – Preference Shares

In addition to ordinary shares, a company may issue other classes of equity such as *preference shares*. These shares are usually senior to the ordinary equity, but junior to bonds, and usually carry a *fixed dividend* such as 5% per annum based on their face value or a fixed amount such as US\$5. This dividend can be *cumulative* or *non-cumulative*. In the case of cumulative shares, if a declared dividend is not paid in one year, then when the next dividends are paid the missed dividend is also paid, i.e. the dividends are rolled up for cumulative preference shares. Missed dividends are not made up in the case of non-cumulative preference shares. It is important to note that, if the dividends are not paid on the preference shares, no dividends can be paid on the ordinary (or common) stock since preference shares are senior to common stock. In the case of cumulative preference shares, all outstanding dividends have to be paid prior to any dividend payment to the common stock holders. Generally preference shares carry either limited voting rights or no voting rights at all. Usually the total amount of preference shares issued is much smaller than the common stock: for example, IBM-A preference shares represent less than 2% of the IBM-issued common stock. The smaller amounts naturally lead to lower trading volumes and less liquidity (i.e. wider bid–offer spreads in quoted prices and smaller trade lot amounts).

Dividend payments by the company are not generally considered a business expense in the way that interest on bonds is tax-deductible/offsettable as they are a distribution of benefits of company ownership. However, some holders (typically companies) of preference shares may benefit from a lower rate of tax on preference dividends compared with the common stock dividend.

I.C.5.2.4 Equity Price Data

Details of trading activity of stocks in the market are distributed widely via electronic and print media. This may be 'real-time' (almost as it happens) or delayed or summary statistics. Daily

summary statistics are produced by journals such as the *Wall Street Journal* and the *Financial Times* and contain information typically like that given in Table I.C.5.3.

Table I.C.5.3: Equity data *Pharmaceuticals & Biotech*

Share	Price	Change	1y High	1y Low	Yield	P/E Ratio	Volume
Glaxo 25p Ord.	1130 xd	+5	1395	980	3.6%	14.7	18,941,000

The volume is the total number of shares bought and sold on the day, and in this case the total volume traded was close to 19 million shares. With a price of £11.30 per share, the value of this daily turnover was approximately (£11.30 × 18,941,000) £214m. This information is usually grouped by market sector, and in this case Glaxo is a member of the Pharmaceutical and Biotech group.

I.C.5.2.5 Market Capitalisation (or ‘Market Cap’)

The market capitalisation value of a listed company is the total amount of issued share capital multiplied by the current share price. This represents the total value or worth of the company. For example, a company with a share price of £10 and total number of shares issued of 100,000,000 would have a market capitalisation of £1 billion. The sum of all the market caps of all the listed shares gives the total stock market value. The market cap is often used as the weighting factor in the calculation of stock market indices such as the Standard and Poor’s (S&P) 500 and the Financial Times Stock Exchange (FTSE) 100.

I.C.5.2.6 Stock Market Indices

Stock market indices such as the S&P 500 and the FTSE 100 are used to measure broad equity market performance and to benchmark investment portfolios. Most indices are weighted by market cap, although simple price-weighted indices do exist (e.g. Nikkei 225, Dow Jones). Price-weighted indices are generally avoided due to the ease with which they can be manipulated by unscrupulous traders. With a simple price-weighted index a price movement in the smallest and most illiquid stock has the same effect on the index as the same move in price of the most liquid and largest company. When derivatives contracts such as futures and options are settled from this type of index, the index value may be pushed up or down relatively easily (and cheaply) by trading in the illiquid stocks. Vastly more profit is made from the derivatives’ settlement value than the cost of manipulating the index. This type of market manipulation is illegal in many jurisdictions. A cap-weighted index is based on the sum of all the cap-weighted prices of the constituent companies. In this case movements in price are scaled by the economic size of the company and so it is much harder (and more expensive) to manipulate its value. In the case of the FTSE 100,

the index contains the largest 100 companies listed on the London Stock Exchange, the ‘largest’ being defined by market cap.

For instance, suppose ABC plc has a market cap of £100m and that the total market value of the top ten stocks is £2000m. Then the weighting of ABC in the ‘top ten index’ will be 5% (i.e. 100/2000). If the mid-price of ABC is £5.00 at close of business, the index calculation is:

$$I = k \times \left(\text{£}5 \times 5\% + \sum_{i=2}^{10} p_i w_i \right) \quad (\text{I.C.5.1})$$

where k is a ‘starting factor’ that sets the index to 1000 on the starting reference day, p_i are the stock prices and w_i the market-cap weights of the other nine stocks in the index.

The FTSE 100 index started in the early 1980s and had a starting reference value of 1000. The index has risen to almost 7000 and fallen back to almost 3000 in the intervening 20 years. The index today does not contain the same 100 stocks as at the start due to changes in market cap (old businesses merging/failing or new companies growing).

Summary statistics can be produced for the index in the same way that they can for a single stock, thus we talk about the market yield or the market P/E ratio, which is usually based on an index such as the FTSE 100 or FTSE All-Share Index. These indices often form the reference underlying price for derivatives contracts such as futures and options. On the FTSE 100 and the S&P 500 there are both exchange-traded and over-the-counter (OTC) derivative transactions.

Some indices such as the DAX 30 are based on ‘total return’, so that they are adjusted to include dividend payments over time. An index including both price movements and dividends is sometimes referred to as an *accumulation index*.

I.C.5.2.7 Equity Valuation

Clearly, one way to establish the value of a company is to analyse the balance sheet and calculate the net book value of its shares. This value is frequently much less than the market price of the stock as it does not take into account future earnings and the value of the company as a going concern. An alternative approach is to value the equity as the present value (PV) of all future dividend payments – this is the so-called ‘dividend discount model’:

$$PV = \sum_{i=1}^{\infty} \frac{D(1+g)^{i-1}}{(1+r)^i} \quad (\text{I.C.5.2})$$

where D is the expected next annual dividend, g is the growth rate of dividend payment and r is the discount rate. The discount rate should represent the rate required by investors to compensate them for both (a) the time value of money and (b) risk of loss. It could be determined using the capital asset pricing model (see Chapter I.A.4), for example.

This dividend discount model can be simplified to the ‘Gordon growth’ model:

$$PV = \frac{D}{r - g} \quad (\text{I.C.5.3})$$

So, for example, with an expected Glaxo dividend of 40p, a dividend growth rate of 2.5% p.a. and a long-term discount rate of 5%, the Gordon growth model would imply a price for Glaxo of £16.00 compared with a market price of approximately £12.00. Clearly this valuation method is critically sensitive to long-term discount rates and the future growth rates of dividends.

Valuation of equity shares is very difficult due to the many sources of uncertainty and the long-term nature of the business enterprise. Ultimately the market mechanism determines the price of equity by matching supply and demand at the ‘market price’. This is underpinned theoretically by techniques such as net book value, P/E ratios and the dividend discount model, as well as more complex corporate and market analysis.

I.C.5.3 Stock Markets and their Participants

The stock market exists to bring together buyers and sellers of equity risk. This facilitates the efficient raising of capital and diversification of risk necessary for capitalism to thrive. When a company needs to raise new capital it may do so by selling shares; when an investor wishes to deploy excess capital to earn a return he or she may do so by buying shares. As simple as this sounds, it requires substantial resources, such as capital and technology, along with regulation to ensure an efficient and fair marketplace in which the participants have confidence. Confidence is an essential component of the market, since, if it considered corrupt or unfair, it will rapidly lose favour with the disadvantaged party and fall into disuse.

I.C.5.3.1 The Main Participants – Firms, Investment Banks and Investors

Companies (or ‘corporations’ or ‘firms’) issue equity to raise capital and diversify their risk among a wider ownership group. This capital raising can occur at the commencement of the business or at some time later when extra funds are required for expansion or to shore up existing activities. The initial issuance of new equity capital by a company is called a *primary issue* and companies access the stock market via a *listing* (or *float*) of their equity securities on the stock market. This listing requires compliance with specific stock market rules and regulations and is usually

undertaken by a specialist financial company (an *intermediary*) such as an *investment bank* or *stockbroker*.

The intermediary will use accountants and lawyers as necessary for compliance with both company law and stock exchange listing requirements (sometimes referred to in the UK as the *yellow book*). In the USA, compliance is with the Securities Exchange Commission (SEC) and the New York Stock Exchange (NYSE). Once a listing can be achieved, details of the new issue will be circulated to *investors* by the financial intermediary, who will be asked to subscribe to (i.e. to *buy*) the issue. Frequently, the deal is 'bought' in its entirety by the financial intermediary, who then takes the risk of selling it on, so guaranteeing the amount of capital raised to the issuing company. Alternatively, the financial intermediary may *underwrite* the issue, so that, if not all the 'paper' can be placed with the investors, the intermediary buys up the surplus 'rump', usually at a slight discount to the issue price. A fee is charged by the intermediary for the work involved in listing and for the risks taken in undertaking a bought deal or underwriting. This fee is usually realised by buying the equity securities at a discount to the expected market price. Fees are negotiable and there is intense competition for 'big-name' issues.

Once stocks have entered the market via a primary issue they trade in the secondary market via financial intermediaries such as investment banks and brokers. Commissions are charged by the intermediaries for arranging and settling secondary market trades. The secondary market activity may be organised in two basic ways:

- *Matched market*. Orders for sale or purchase with amounts are entered into a system or 'order book' with a *limit price* at which the investor is happy to sell or buy. The system matches trades at the best price that is acceptable to both counterparts and 'crosses' the trade at the matched price. This system can lead to wild swings in prices and periods of illiquidity.
- *Market maker*. Financial intermediaries make two-way prices (bid and offer prices in market lot sizes) which can be *hit* (i.e. to hit the bid is to sell) or *taken* (to take the offer is to buy) by investors, leaving the market maker with a risk position that he or she must manage. The market maker therefore uses their own capital to create a more liquid market and to damp wild price swings.

Often, secondary market trading is a mixture of these two methods, with larger stocks trading via a market-maker approach and smaller, less liquid stocks trading on a matched-market basis.

I.C.5.3.2 Market Mechanics

The market requires secure communication between qualified participants. In earlier times the right people met in a designated secure room or place at an agreed time and traded directly with each other. Today the market is largely electronic or telephonic so that details of stock prices and

trade amounts are communicated between financial intermediaries and transactions agreed either electronically or by phone. Investors deal through a broker of their choice, who may also 'hold' the equity securities on their behalf in *custody* systems. These allow rapid movement of stocks electronically to settle trades. In the most advanced systems, settlement of trades can be achieved almost in real time. Once a trade is agreed, secure electronic messages can be generated and sent, which debit the buyer's cash account and credit his securities account with the purchased stock, while simultaneously doing the reverse to the seller's accounts. In less developed markets, the settlement process involves the movement of physical paper securities, with attendant delays and risk of a failed trade.

When referring to trading and settlement, brokers typically refer to T+1 or T+3 settlement. In this terminology a trade originated on 'trade day' (T) will be settled (i.e. stock transferred and payment made) on a settlement day sometime later. So T+3 settlement denotes a three-day delay between trade day and settlement day. This is sometimes called 'rolling settlement' as opposed to the old system of an account period where all trades made within an account were settled at the same time. Historically, the UK stock market had a three-week account period, which has now been changed to a rolling settlement of T+3. Ultimately the goal is to reduce this to T+1 or even same-day settlement. Shorter settlement periods are very useful in reducing the risk of default by a counterparty prior to settlement and in helping to minimise the effects of 'out-trades' or dealing mistakes, as these are spotted earlier and rectified before the stock price has moved too far.

I.C.5.4 The Primary Market – IPOs and Private Placements

Initial public offerings (IPOs) is the name given to a formerly privately owned company selling equity securities to third-party investors for the first time, sometimes known as *floating on the market* or *listing*. *Seasoned new issues (SNIs)* is the name given to companies issuing securities after they have floated. Both IPOs and SNIs may be made via a *public offering*, which makes the securities available to the general investor population, or via a *private placement*, in which the issue is placed directly with a few specially chosen investors and is not widely traded after issue. IPOs and SNIs are examples of the *primary securities market* or *new issues market*.

I.C.5.4.1 Basic Primary Market Process

We will use US market practice as our example, but the approach is broadly similar to the approach taken in other major markets such as the UK or Eurozone.

The firm wishing to float contacts a number of investment banks to negotiate terms with regard to an IPO. Terms include fees and costs in addition to marketing strategy and experience. The firm chooses a bank to be the lead player in the IPO; sometimes this will involve underwriting

the new issue if they fail to place all the paper at launch. Often an underwriting syndicate, headed by the lead manager, will be formed to share the risk and broaden the distribution of the securities. The lead manager advises on terms and pricing of the IPO. A preliminary notice is filed with the SEC (one of the listing authorities in the USA), giving basic terms and details of the issuing company. This is a preliminary prospectus, which has to be finalised and approved by the SEC prior to its becoming the IPO prospectus. This is then used to market the securities to the investors and the issue price is fixed. This process may take several weeks or longer.

If the issue is *fully underwritten*, the entire issue is bought at launch by the underwriting syndicate at a discount to issue price and then distributed to investors; this is also known as a ‘firm commitment’. Compensation for bearing this risk is the size of the discount to issue price and the tightness of the pricing. This method is more common for bond issues of high-quality borrowers than for equities.

If the issue goes ahead on a ‘*best efforts*’ basis, the price risk remains with the floating company and not with the bank. The bank collects a fee for arranging the IPO and a sales commission for stock sold (at the height of the dotcom boom, it was not unusual for banks to be paid in stock rather than cash). This method is the more common approach for an IPO of common stock.

I.C.5.4.2 Initial Public Offerings

The lead manager is responsible for marketing the new issue once the SEC has accepted the issues registration document and preliminary prospectus. This marketing may take the form of a ‘road show’ to investors, having two main aims:

- informing investors of the floating company and its activities, emphasising its attraction as an investment; and
- sounding out the investors as to likely price levels at which they will purchase the securities at launch.

Talking to investors and getting them to commit to purchase securities at launch is called *book building*, and this process allows fine tuning of the offer price. Strong early commitment by investors is usually rewarded with a large allocation of shares and possibly even a discount. This can lead to substantial underpricing and a large jump in price at issue. At the height of the dotcom bubble, one share closed up nearly 700% on issue day!

Typically, a lead manager will charge around 7% in fees for an IPO, but this does not include any discount to the market price on issue, which as we have seen can be very substantial. In general, IPOs tend to be ‘cheap’ on issue, but not in every case are all the securities sold at issue price. On occasion stock is left with the underwriters. Longer-term studies have shown that many IPOs,

particularly those coming from the dotcom era, have proved to be relatively poor long-term investments. Money was made by those participating in the float and then selling into the market demand that followed very shortly afterwards.

There have been initiatives to move away from the investment-bank-led IPO due to the large fee involved and the potentially substantial underpricing. These have included internet book-building exercises and do-it-yourself IPOs. So far they have had limited success and have focused on the smaller end of the market. At present Wall Street still dominates the IPO business and enjoys relatively generous fees in the process.

I.C.5.4.3 Private Placements

Private placements of common stock are much cheaper than IPOs since the entire issue of securities is sold to a small group of investors under rule SEC 144A (in the USA), which permits a simpler (and therefore cheaper) registration and listing process. The need for extensive road shows is obviated. However, it is difficult to place large issues this way due to the limited risk appetite of a small investor group. Furthermore, these issues tend not to trade in the secondary market, making it difficult for the investor to liquidate his position at short notice. This in turn means that investors in private placements demand a discount price for bearing this additional risk.

I.C.5.5 The Secondary Market – the Exchange versus OTC Market

The secondary market consists of the buying and selling of already issued securities and represents by far the largest volume of activity by value on a day-to-day basis. This activity is effectively investor-to-investor trading via a financial intermediary (in the case of on-exchange and OTC activity) and directly with each other on a peer-to-peer basis.

I.C.5.5.1 The Exchange

Usually each developed country has at least one national stock exchange. In the USA there are two major ones – the American Stock Exchange (AMEX) and the NYSE – and several regional exchanges dealing in smaller local companies. Only members of the exchange are allowed to trade on it, and the membership is called a *seat*. Seats are owned by brokers and banks who in turn deal with their own clients (the investors). Members of the exchange charge a commission for executing trades on the exchange on behalf of their clients. The NYSE has approximately 3000 members and trades in about 3300 common and preference stocks, which represent the vast majority of large and medium-sized corporations in the USA. In order to be listed on the

NYSE, in addition to SEC registration, the company needs to meet the minimum requirements as to size and profitability; the market cap requirement is more than \$60m.

The basic trading mechanism on an exchange is illustrated for the NYSE as an example (other exchanges may vary somewhat):

- an investor places an order with a broker (who owns a seat on the exchange);
- the order is passed to the firm's commission broker on the floor of the exchange;
- he approaches the *specialist*, who is responsible for market making (making two-way prices and managing the order flow) in that particular stock (on the NYSE there is only one specialist per stock);
- the order is placed and dealt; and
- confirmation of the 'fill' flows back to the investor via the broker chain.

This process may be partly physical (i.e. real people in a real room) or electronic via trading monitors (e.g. the Stock Exchange Trading System (SETS) in London, or the SuperDOT system in New York). In some markets the role of broker and market maker can be carried out by one firm and there may be multiple market makers in one stock.

There are essentially two types of order for buying or selling stock on exchange:

- *Market order* – deal the stock at the current market price and size. If the order size is larger than the quoted size (e.g. 55.20 bid/55.25 offer in 1000 shares) then the order is executed at multiple successive prices until it is filled.
- *Limit order* – deal the whole order at a pre-fixed price. Variations on this include 'fill or kill', which means that the entire order must be filled immediately at this price in one go or not at all.

The market maker ('specialist' in the USA) in each stock has a responsibility for maintaining the market in that stock (monitored by the exchange authority). Many transactions, however, will actually be 'crosses' from one broker (buying) to another broker (selling) at a price within the market makers' bid-offer spread. Frequently, very large orders called 'blocks' are traded. Some brokers specialise in taking the other side of such transactions at discounted prices to provide liquidity and to profit from (hopefully) unwinding the large position over time.

I.C.5.5.2 The Over-the-Counter Market

Transactions on exchanges all go via the central market maker(s) or specialists. In the OTC market, deals are done directly between broker/dealers who make two-way prices to each other in the stocks that they trade. Without a central market maker this means that the broker/dealer initiating the transaction has to search for the best price for the deal from a large number of potential counterparts. The North American Securities Dealers Automated Quotation

(NASDAQ) in the USA is an OTC market and brokers/dealers display their quotes via the electronic system, but must actually contact the dealer directly to obtain a firm quote and deal. A trading system called the Small Order Execution Service (SOES) exists alongside the NASDAQ and fills small trades at the stated ('screen') price. NASDAQ is working to upgrade this system with its SuperMontage development. In the UK, the Stock Exchange Automated Quotation System (SEAQ) and SETS perform similar, but not identical, roles.

By their nature OTC markets are diffuse and non-centralised and are therefore ideal for electronic information and trading platforms. This non-centralisation, however, sometimes means that deals are not done at the best market price (sometimes called *trading through*) since not all deals go through a central market maker or specialist. Some may slip through unnoticed by a better buyer or seller (i.e. someone prepared to buy at a higher price or sell at a lower price than that actually dealt). It also makes OTC markets potentially more difficult to monitor and regulate and they may not provide true *price discovery* (i.e. making sure that *all* potential participants have the opportunity to quote), which is a feature of a centralised market such as the exchange.

I.C.5.6 Trading Costs

The costs of buying and selling common stock are a combination of explicit costs, such as commissions and brokerage fees, and more hidden costs, such as the width of the bid-offer spread. There are also market impact costs, when the size of the transaction is sufficiently large that executing it moves the price away from the indicated quote. Total trading costs will therefore vary by market and indeed by stock, and may increase or decrease over time depending on market conditions. Trading costs are significant, not only from the point of view of reducing total return, but also in determining the viability of arbitrage trades such as stock-index/future arbitrage.

I.C.5.6.1 Commissions

The commission paid to brokers is normally negotiable and will depend on the size and volume of trades to be placed via the broker and the level of service expected. This can vary from an execution-only service for large volumes of large-value trades to a bespoke, 'full-service' fund-management process for a small investor, with information, detailed reports and analysis. The latter is clearly more expensive to provide. Some exchanges/markets insist on minimum commission rates to safeguard the smaller broker, but some larger firms may find a way to rebate some of this to their customers, lowering the real cost. Typically, the execution-only broker charges a fixed fee. For orders placed via the Internet (the cheapest method) this can be as low as US\$10–20 or alternatively a smaller flat fee plus a small cost per share (e.g. 2 cents per share). Full-service brokers may charge as much as US\$300 for the same trade placed by telephone to one of their trading assistants.

I.C.5.6.2 Bid–Offer Spread

A major cost difference between on-exchange and OTC deal execution is that on the former many trades will be crossed between brokers inside the indicated market maker bid–offer spread. This is called *price improvement* because the actual deal price is below the initial quoted offer or above the quoted bid when struck – i.e. it is dealt ‘inside’ the spread and is therefore an improvement on the quoted price – whereas, on the OTC market, the client will pay or receive the dealer’s quoted price and thus always be subject to the full bid–offer spread on any ‘round trip’ (buying and then later selling) in the stock. The client of course just ‘sees’ the price dealt at and may never explicitly recognise this cost. Typical bid–offer spreads in large liquid common stocks are of the order of 0.5–1.0%, and so Glaxo with a mid price of £12.00 may have a quoted bid–offer of £11.97–12.03 in, say, 5000 shares.

Less liquid stocks and less busy market makers and exchanges tend to increase the market bid–offer price in a stock and hence the costs of trading. In OTC markets there is always the risk that deals are being routed to dealers who do not offer the best prices, and it is difficult to ensure ‘best execution’. Estimating the real cost of the bid–offer spread is clearly not simple.

I.C.5.6.3 Market Impact

When a trade is executed it represents new information in the market and the market price reacts. Buying stock should drive up the market price, all else being equal. If the deal is large and the stock is illiquid, the actual trade execution price will be higher than the indicated bid–offer price due to its ‘market impact’. Market impact is a function of the ‘depth’ (number of potential buyers and sellers outside of the bid–offer) of the market at the time of execution and will vary over time and by stock. Estimating market impact is usually the result of studies of market price reactions with trade size over a suitably large number of market conditions, usually indicated by share turnover or futures volume. Estimating market impact, as with real bid–offer costs, is potentially complex and requires data and modelling.

I.C.5.7 Buying on Margin

Essentially, buying stock on margin consists of taking a loan from the broker (a *broker call loan*) to buy more stock than his own funds would allow. The investor *leverages* his position in a stock through a combination of his own and borrowed funds. Once purchased, the stock remains with the broker as collateral for the loan. The investor has to pay interest on the loan and a fee or commission to the broker for the arrangement.

I.C.5.7.1 Leverage

Leverage is the use of borrowed funds to allow an investor to take a larger risk position than he would ordinarily be able to do with his own funds. An investor with \$10,000 who buys a position in securities worth \$20,000 is leveraged two times and has borrowed \$10,000 in order to achieve this. In this case the investor has \$10,000 of his own equity and \$10,000 of broker call loan in his margin account. Synthetic leverage can be obtained by using derivative products such as futures and options. For example, the payment of option premium will give the investor the right to buy or sell a much greater value of stock than would be possible with a direct transaction.

Some regulatory authorities set limits on the amount of leverage that can be offered by margin trading. In the USA, the Federal Reserve set a limit of 2× so that only up to 50% of the money invested can be borrowed. This limit is varied from time to time, dependent on the perceived risks of too much or too little gearing in the system. In other markets the degree of leverage offered is at the discretion of the broker/dealer but subject to regulatory oversight/control via the firm's capital adequacy. If the regulator considers that a firm is offering too high a leverage to its clients, thereby increasing its risk of default, the regulatory supervisor may insist on an increase of qualifying capital set aside to cover this risk.

I.C.5.7.2 Percentage Margin and Maintenance Margin

Once the loan has been agreed, subject to the maximum leverage not being exceeded, the money is invested in the stock. The value of this stock may rise or fall and this will affect the amount of the investor's 'equity' or 'own funds' in the position.

For instance, suppose an investor uses £5000 of his own money and a loan of £5000 to buy £10,000 of stock. Then his leverage is two times and his percentage initial margin (equity/stock value) is 50%; if the value of the stock falls to £8000, the investor has £3000 of equity or own funds remaining (he has lost £2000 on a mark-to-market basis) and a percentage margin of $\frac{£3000}{£8000} = 37.5\%$. Clearly the collateral is still sufficient to cover the loan of £5000.

If the value of the stock falls sufficiently far that the investor's equity is close to zero, the broker makes a *maintenance margin call*. That is, she requires that the investor *top up* his account so that his percentage margin is above a minimum value (say, 10%). This is known variously as the *trigger* or *margin call rate*. If the client did not respond immediately (in cash or by pledging securities) and pay his margin call, the broker could liquidate or close out the position in the stock to protect her collateral on the loan. Clearly the more volatile the stock price, the more urgent the margin call can become and sensibly the higher the trigger rate should be set. Typically, investors would

have, at most, a few days to meet a margin call and in times of extreme volatility the broker/dealer generally has the right to close out the client without notice.

I.C.5.7.3 Why Trade on Margin?

By borrowing and buying a larger stock position than the investor would ordinarily be able to, he can create a leveraged risk position. Clearly, a necessary requirement for the investor is that the expected rate of return on the investment be greater than the cost of the loan. We consider a stock that rises 15%, 0% and minus 15% over one year and two investors A (who is unleveraged) and B (who is two times leveraged and borrows at 10%). We compare their risk–return ratio in Table I.C.5.4.

Table I.C.5.4: Margin trade – risk and return

Stock Price	Investor A	Investor B
+15%	+15%	+20%
0%	0%	–10%
–15%	–15%	–40%

Clearly, the leverage position increases both upside and downside, but the relatively high interest costs (10%) drag down the returns for investor B. Margin trading works best for investors who are able to use it for very short-term positions (e.g. a week) where a rapid movement in the stock price may yield an annualised return very much higher than the interest costs.

I.C.5.8 Short Sales and Stock Borrowing Costs

Short-selling is the process of selling a security that the investor does not own with the intention of buying it back more cheaply later to make a profit. The *short sale* is the method by which an investor can speculate on the fall in share prices rather than their rise. In order to sell short it is necessary to *borrow stock* for delivery in the initial sale trade. Then, when the position is to be closed out, the shorted shares are bought in the market and returned to the counterparty who lent the stock. This is called *covering the short*. Clearly, the lender of the stock demands a fee for this service, and this is known as the *stock-borrowing cost* or *repo* cost. The word ‘repo’ or ‘RP’ is an abbreviation of ‘sale and repurchase agreement’, which is a more common form of managing short positions in the fixed-income market. In that market a repo involves an agreement to sell and buy back rather than borrow and return securities.

I.C.5.8.1 Short Sale

In some markets, notably the USA, there are restrictions on when a short sale can occur. The so-called *up-tick* rule prevents a short sale unless the last price move in the stock was positive. This rule is designed to limit the volatility of market swings. Further rules prevent brokers/dealers

from investing the proceeds of the short sale in other positions, thus limiting the amount of leverage that can be generated this way. In other markets (e.g. the UK) the up-tick rule does not apply and overall leverage of the firm is controlled via capital adequacy. Under the Capital Adequacy Directive rules, firms calculate their potential loss exposure to investors using either a simple rules-based approach or a more complex risk model (which simulates movements in the value of collateral) and then allocate capital against this requirement. As a firm's capital is finite, this places an upper limit on the total risk the firm can take and in turn the degree of leverage it can offer to clients. In some markets short selling is restricted or may not be allowed at all from time to time in an attempt to protect the market. This occurred in markets in the Far East such as Malaysia and Thailand during the turmoil of the late 1990s.

An investor selling short via a broker/dealer is required to post margin as in the margin-trading example above. This is due to the fact that a rise in the stock price will leave the investor exposed to a mark-to-market loss which the broker/dealer will need to cover. Hence the usual leverage, initial margin and maintenance margin considerations apply in short selling as well as margin trading.

I.C.5.8.2 Stock Borrowing

Typically, stocks are lent by brokers/dealers from securities that are pledged or held in custody on behalf of their clients. Large investors who hold their own stocks (e.g. insurance companies) may lend directly in the market. The loan may be at *call*, which means that it may be terminated at any time by the lender (which is the market standard), or a *term* loan for a predefined period (e.g. one month). Note that the loan is for a specific number of shares, not for a specific value, since any change in share price will change the value of the loan. Stock borrowing is normally a secured loan activity so that, when shares are lent, cash or more likely securities are pledged in return as collateral. Typical stock borrow/loan fees in large European stocks are 30 basis points (0.3%) per annum. If a particular stock is in short supply to lend – e.g. when the market is very bearish on the company and many investors are 'short' – this rate may rise to 20% or 30% p.a. When stocks are borrowed over a dividend payment date the stock has to be returned to the lender along with a payment for the dividend paid while on loan.

Usually the arrangements for stock borrowing, collateral and fees are handled directly by the broker/dealer for the investor. All the investor sees are the net interest costs on his account and the margin calls. In certain markets (notably the Far East) stock borrowing is restricted or not allowed. This has a profound effect on trading activity. For, example, arbitrage strategies that involve short-selling stock against a long position in convertible bonds may not be possible.

I.C.5.9 Exchange-Traded Derivatives on Stocks

The pricing of options and futures, along with their risk characteristics, is discussed in Chapters I.A.7 and I.A.8. We will focus here on describing some of the stock derivatives available and how their markets operate.

I.C.5.9.1 Single Stock and Index Options

Tens of derivatives exchanges exist across the world, but we will focus on two of the major stock option exchanges. These are the Chicago Board Options Exchange (CBOE) in the USA and London International Financial Futures Exchange (Euronext.LIFFE) in the UK. These markets offer a range of derivative products that have common stocks as their underlying asset, either single stocks or stock indices, which in effect behave like baskets of stocks.

On the CBOE, single-stock options exist on approximately 500 individual companies, most of which are in the S&P 500 index. These stock options allow the holder the right but not the obligation to buy (call option) or sell (put option) 100 shares (which is a normal lot size) of the underlying stock. On exercise (these options are American style, so may be exercised at any time up to expiry) the contract is settled by physical delivery of shares rather than in cash. In addition to single-stock options, there are options on stock indices such as the S&P 500 – these options are cash-settled – and also options on the S&P 500 futures contract. On LIFFE, single-stock options exist on approximately 90 individual companies, most of which are in the FTSE 100, and there are options on the FTSE 100 stock index itself, which are cash-settled.

I.C.5.9.2 Expiration Dates

CBOE stock options expire at 10.59 p.m. (Central Time) on the Saturday following the third Friday of the expiration month. There are January, February and March cycles. The January cycle is January, April, July and October, and the other cycles lag by one and two months respectively. This means that generally the maximum maturity of the option is nine months. Longer-dated options – long-term equity anticipation securities (LEAPs) – on stocks do exist, with expiries in January out to three years. There is a similar three-month cycle for the Euronext.LIFFE stock options.

I.C.5.9.3 Strike Prices

The strike prices are chosen when a new expiry series is introduced, normally with \$2.50, \$5.00 and \$10.00 spacing on either side of the current spot price to give five option strikes in both puts and calls. If the stock price moves outside of the range of existing strikes, a new series is introduced at the new spot strike price.

I.C.5.9.4 Flex Options

These options are agreed individually between brokers with an OTC-like flexibility in expiry and strike price, along with a choice of American or European exercise. They are an attempt to win on-exchange business (with the attendant safety from on-exchange margin arrangements) from the OTC options market. There is an exchange-specified minimum size.

I.C.5.9.5 Dividends and Corporate Actions

CBOE exchange options are not adjusted for cash dividends, so any payments that reduce the stock price are not adjusted in the strike price of the options. If there are corporate actions, such as a stock split (e.g. two new shares for one old one), these are adjusted for in the strike price of the exchange options and the number of shares involved. In this case (i.e. a two-for-one stock split) each strike price would be halved and the number of shares deliverable doubled.

I.C.5.9.6 Position Limits

There are maximum numbers of contracts that can be held by individual investors or groups of investors working in concert. This is to prevent the market being cornered and manipulated to the detriment of other market participants. On the CBOE the largest limit is 75,000 contracts for the largest single stocks.

I.C.5.9.7 Trading

Both the CBOE and LIFFE use market makers to facilitate trading who are committed to maximum bid–offer spreads to aid market liquidity. The markets these days are largely electronic, having evolved from pit trading in the 1980s and 1990s. The trading process is via a broker/dealer through a market maker. Settlement and margin management is via the exchange clearing house. On the CBOE this is the Options Clearing Corporation, whose members are the clearing brokers on the exchange. For further details and contract specification, contact the CBOE or Euronext.LIFFE via their websites (see ‘References’).

I.C.5.10 Summary

In this chapter we have looked at the structure and organisation of the stock market and the common stocks that trade on it. We have considered the characteristics of common stock and the role of the participants in the market. We have studied in detail how new issues come to the primary market (or float) via IPOs or private placements, and how they then trade among investors via brokers/dealers in the secondary market. We have looked in detail at the costs of trading, including short selling and trading on margin. We have also seen how exchange-traded derivatives are organised to trade stock risk via futures and options. For further information and details see ‘References’.

References

Bodie, Z, Kane, A, and Marcus, A J (2002), *Investments* (Boston: McGraw-Hill/Irwin).

Chicago Board Options Exchange – Stock Option Details,
<http://www.cboe.com/OptProd/EquityOptions.asp>.

Hull, J C (2003), *Options, Futures and Other Derivatives* (New York: Prentice Hall).

London International Financial Future Exchange – Stock Option Details,
<http://www.liffe.com/products/equities/index.htm>.

London Stock Exchange – Listing Requirements, <http://www.londonstockexchange.com/>.

New York Stock Exchange – Listing Requirements, <http://www.nyse.com/>.

Reverre, S (2001), *The Complete Arbitrage Desk Book* (New York: McGraw-Hill).

I.C.6 The Futures Markets

Canadian Securities Institute¹

I.C.6.1 Introduction

This chapter provides an introduction to futures markets where exchange-traded, forward-based derivatives are traded. Forward-based derivatives represent contracts made between two parties that require some specific action at a later date. Most often, this action takes the form of delivery of some underlying asset and payment for the asset. All forward-based contracts have a buyer and a seller, a maturity or expiration date, and a formula for exchanging payments set up when the contract is initiated that takes effect at some later date. Apart from a performance bond, no up-front payment is required. All forwards are in effect zero-sum games. The buyer's gain will be the seller's loss and vice versa. The gain and loss will always have a linear relationship with the price of the underlying interest. It should be noted that all forwards facilitate the use of leverage.

A forward-based derivative can trade on an exchange or over the counter (OTC). When it is traded on an exchange, it is referred to as a *futures contract*. There are two general types of futures contracts. Contracts that have a financial asset as their underlying interest are referred to as *financial futures*. These would include interest-rate, currency and equity futures. Contracts that are based on a physical or 'hard' asset are generally referred to as *commodity futures* contracts. Examples of commodity futures are gold, soybeans and crude oil. A detailed description of spot and future commodity markets is given in Chapter I.C.7.

I.C.6.2 History of Forward-Based Derivatives and Futures Markets

Forward-based derivatives have been around for centuries. Initially, they were largely based on agricultural products. Volatile financial markets in the 1970s led to the concept of forwards being applied to financial products such as stocks, bonds and currencies. In 1968, approximately 15 million futures contracts were traded world-wide, predominantly based on agricultural commodities on US exchanges. In 2003, about 3.0 billion futures contracts and 5.1 billion options on futures contracts were traded.² The vast majority of these contracts were based on non-agricultural assets on 48 exchanges in 25 different countries. For the ninth straight year, volume was higher on exchanges outside the US than on US exchanges. Recently, exchanges in

¹ Canadian Securities Institute, Toronto.

² *Futures Industry*, March–April 2004.

Mexico, China and India have all increased their world rankings and are now included in the top 15 exchanges world-wide (Table I.C.6.1).

Table I.C.6.1: Top 15 futures exchanges

2003 Rank	2002 Rank	Exchange	Volume of contracts traded in 2003*
1	1	Eurex	668,650,028
2	2	Chicago Mercantile Exchange	530,989,007
3	3	Chicago Board of Trade	373,669,290
4	4	Euronext	267,822,143
5	7	Mexican Derivatives Exchange	173,820,944
6	6	BM&F (Brazil)	113,895,061
7	5	New York Mercantile Exchange	111,789,658
8	8	Tokyo Commodity Exchange	87,252,219
9	10	Dalian Commodity Exchange (China)	74,973,493
10	9	London Metal Exchange	68,570,154
11	11	Korean Stock Exchange	62,204,783
12	12	Sydney Futures Exchange	41,831,862
13	22	Shanghai Futures Exchange	40,079,750
14	25	National Stock Exchange of India	36,141,561
15	13	Singapore Exchange	35,356,776

*Volume figures exclude options on futures.

Source: *Futures Industry*, March–April 2004

The agricultural industry was largely responsible for launching forward trading as producers (farmers) and consumers (millers) sought to minimise price uncertainty. Although agricultural prices fluctuate with supply and demand like most other prices, seasonality and weather conditions tend to make these fluctuations more severe and unpredictable. For example, an unusually large harvest can overwhelm markets with excess supply, causing prices to fall. Similarly, as supplies are drawn down after harvest, shortages and escalating prices can result. The concept of forward buying and selling was developed to help producers and consumers protect themselves against seasonal price fluctuations.

The Japanese were the first to introduce forward trading in the 1600s with rice forwards. In North America, the grain industry was the first to embrace forward-based contracts. Initially, contracts were developed in which a buyer and seller agreed privately, in advance, to the terms of a sale that would be consummated when the goods *arrived*. These agreements, known as *to-arrive*

contracts, had their origin in the Liverpool cotton trade in the late 1700s. In the beginning, buyers and sellers met in the street to conduct business, but as volumes grew, a more permanent marketplace was sought.

Although the to-arrive contracts helped smooth out seasonal boom and bust cycles, they were not a perfect solution. Disputes often arose at delivery over the terms of the contracts, and the threat of default was always present. The private nature of the contracts meant pricing information was limited. The buyer and seller in a particular deal were generally unaware of prices from other contracts and therefore would have difficulty determining the current market price. Another problem concerned contract resale. Early contracts were not transferable. Even when they became transferable, it was difficult for a buyer or seller to find a third party willing to accept the risk.

Many of these problems, with what were essentially OTC forwards, were resolved with the introduction of exchange-traded forwards which became known as futures contracts. The Chicago Board of Trade (CBOT) became North America's first organised futures market in 1848 when buyers and sellers moved off the street and into the exchange. However, it was not until the 1860s that the innovative concept of standardised contract terms was introduced.

Listed futures contracts were standardised in terms of size, quality, grade, and time and place of delivery. Standardisation, together with the requirement that all trading take place in a single location via the *open outcry system*, facilitated accurate and immediate price dissemination. Soon, a margining system was developed to guarantee the financial integrity of each contract.

The development of futures trading attracted not only the merchants involved in the grain trade, but also individuals who were only interested in the market for its profit possibilities. The influx of *speculators* greatly improved liquidity, which helped enhance market efficiency. Liquid futures markets helped eliminate the risk of being unable to resell and also helped to minimise wide price fluctuations.

The success of exchange-traded grain contracts led to tremendous growth in new futures contracts and new exchanges. Cotton, lumber, livestock, coffee and orange juice futures were eventually followed by industrial and precious metals. In the early 1970s, the first foreign currency contracts were developed, followed by contracts on debt instruments starting with the Government National Mortgage Association futures (GNMA or 'Ginnie Mae'). In the early 1980s, the next generation of futures complexes, stock index futures, was initiated with the

introduction of the Value Line and S&P 500 index contracts. Energy-based futures began trading in the mid-1980s.

It is evident that futures contracts developed to solve some of the problems associated with OTC forward agreements. Their growth was so rapid that, not too long after their inception, futures markets became the predominant market for transacting forward-based contracts. This predominance became even more pronounced with the inception of precious metal and then financial futures contracts. From non-existence in the early 1970s, financial futures (interest-rate, currency and stock index contracts) now account for approximately 80% of all futures trading. They have been primarily responsible for the almost exponential growth in overall futures volumes. Table I.C.6.2 highlights the relative importance of financial futures in world markets as of 2003.

Table I.C.6.2: Global futures and options volume by sector

	2003 Volume in millions	% change on 2002
Equity indices	3,960.87	41.91%
Interest rate	1,881.27	27.25%
Individual equities	1,558.52	15.05%
Agricultural commodities	261.15	30.98%
Energy products	217.56	3.91%
Non-precious metals	90.39	26.29%
Foreign currency/index	77.85	28.53%
Precious metals	64.46	25.75%
Other	0.66	-17.14%
Total volume	8,112.73	30.49%

Source: *Futures Industry*, March–April 2004

I.C.6.3 Futures Contracts and Markets

A futures contract is an agreement between two parties to buy or sell an asset at some future point in time at a predetermined price. This section describes the characteristics and mechanisms that are common to all futures contracts, and the highly organised and structured markets in which these contracts are traded.

I.C.6.3.1 General Characteristics of Futures Contracts and Markets

Since futures contracts trade on an exchange, all futures contracts are *standardised* in terms of their size, grade, and time and place of delivery. Other features of futures contracts that are standardised include their trading hours, minimum price fluctuations and, for contracts that have them, maximum daily price limits. All contract terms, except price, are defined by the exchange on which they trade. This standardisation can have an impact on hedging, as delivery dates and terms are not flexible. Table I.C.6.3 provides an example of the standard specifications of a typical futures contract – flaxseed futures that trade on the Winnipeg Commodity Exchange.

Table I.C.6.3: Contract specifications of Canadian flaxseed futures

Trading unit:	20 metric tonnes
Minimum tick size:	C\$0.10 per metric tonne (C\$2 per contract)
Daily price limit:	C\$30 per metric tonne
Delivery months:	January, March, May, July, September and November
Trading hours:	9:30 a.m. to 1:15 p.m. (Central Time)
Delivery point:	Areas in south-eastern Manitoba at par; areas in south-western Manitoba and south-eastern Saskatchewan at a C\$5 pr metric tonne discount; areas in south-central Saskatchewan at a C\$10 per metric tonne discount
Deliverable grades:	#1 Canada western (CW) flaxseed at par or #2 CW flaxseed at a C\$2 per metric tonne discount

The *trading unit* describes the number of units that underlie the futures contract. This is the amount per contract that must be delivered or accepted for delivery if the contract is held to the delivery month. For instance, in the example in Table I.C.6.3, if the current flaxseed price was \$400 per tonne, the value of the contract would be \$8000 ($\400×20 tonnes). At delivery, the seller would deliver \$8000 worth of flaxseed that the buyer would have to pay cash for.

The minimum *tick size* represents the smallest price increment the futures contract can move up or down. In the case of flaxseed, it is 10 cents per tonne. If the current price of an October flaxseed futures contract is \$400 per tonne, the next trade could take place at a price of either \$400, \$400.10 or greater, or \$399.90 or less. The 10 cent per tonne increment translates to \$2.00 per contract (20 tonnes \times 10 cents).

Exchanges set *limits* on the amount by which most futures can move, either up or down, during one day’s trading session. If the price moves down by an amount equal to the daily limit, the

contract is said to be limit down. If it reaches the upper limit then it is said to be limit up. The limits are designed to calm market panic, and to give market participants time to absorb new information that may have been disseminated.

Example I.C.6.1: Daily Trading Limits

A severe drought on the Canadian prairies resulted in volatile trading in the flaxseed futures contract that trades on the Winnipeg Commodity Exchange. After a settlement price of C\$450 per metric tonne on the previous trading day, July flaxseed futures moved up by the daily limit of C\$30 to C\$480. The C\$30 limit prohibits any trading from taking place over C\$30 above or under C\$30 below the previous day's settlement price of C\$450. In this particular case, no trades may take place above C\$480 or below C\$420.

When a futures price moves by its daily limit, there still may be some trading at the limit price. Most often, however, trading comes to a complete halt as *bids* (the highest price at which someone is willing to buy) or *offers* (the lowest price at which someone is willing to sell) are non-existent when the market moves limit down or up, respectively. This kind of situation can be very dangerous for traders holding losing long or short positions because they are unable to liquidate. If the limit situation lasts for several days, huge losses can result.

Partially in recognition of this risk, most exchanges have adopted procedures to deal with limit moves. One procedure expands price limits after a few days of limit moves. *Expanded limits*, for example, may widen out to 150% of regular limits so as to give traders holding losing long or short positions a greater chance to liquidate. Another procedure removes limits entirely for futures contracts trading in their delivery month. Finally, some exchanges have abolished limits on some contracts altogether.

Futures contracts are *settled daily*. Profits are credited daily to accounts that have winning positions, and losses debited daily to accounts that have losing positions. The size of the daily amount depends on the relationship between the current futures price and the initial entry price. If the futures price is higher, the holder of the long position receives a payment from the short for an amount equal to the difference. If the futures price is lower, the holder of the short position receives a payment from the long for an amount equal to the difference.

A futures contract only gains or loses value as the futures price changes. The payoff from a position in a futures contract is *linear* and, because of margining and daily mark-to-market, there may be significant cash flows associated with futures contracts. Cash flows can be positive or

negative and, if not properly anticipated, can affect a party's ability to effectively use futures as a hedging tool.

The *delivery months* are also set by the exchange. In addition, the exchanges set specific deadline days for when trading in a contract ceases and for when the delivery period begins and ends. In the case of flaxseed, the last trading day for a particular delivery month is always the eighth last business day of the delivery month. The first delivery day is always the first business day of the delivery month.

The exchange also sets the deliverable grade (the quality of an asset that will be accepted for delivery in terms of grade, weight or other characteristics), and other alternative grades that are acceptable for delivery. The deliverable grade for flaxseed is #1 CW flaxseed. However, #2 CW will also be accepted, albeit at a \$2 per tonne discount off the final settlement price.

A *clearing association* stands between the parties to a futures contract. As a result, counterparties' identities are irrelevant. Companies A and B, which may have equal and opposite positions in futures contracts, can easily terminate their respective futures contracts at anytime following onset up to contract expiration by what is referred to as an *offsetting transaction* (see Section I.C.6.3.2). Company A could independently sell and company B independently buy the contract in the secondary market, which would have the effect of liquidating their respective positions and have no dependence upon each other because of the clearing association: if A or B defaulted on their obligations, the exchange would assume the obligations of the defaulted party. Section I.C.6.5.3 gives more details about the activities of clearing associations.

The financial integrity of the futures markets is protected by requiring that each party to a contract to post a performance bond, which is called the *margin*. Through a daily *marking-to-market* process with corresponding transfers of margin, each party to a contract is assured of the other party's performance. The initial value of a futures contract to both buyer and seller is zero, but initial margins, which are discussed in Section I.C.6.6, are not.

Finally, futures markets are *regulated* by governmental agencies and self-regulatory organisations. Regulations are very specific and detailed. Before any futures contract can be listed for trading it must be approved by regulatory authorities. For example, in the USA all new contracts must be approved by the Commodity Futures Trading Commission; in Singapore all new contracts must be approved by the Monetary Authority of Singapore.

I.C.6.3.2 Settlement of Futures Contracts

This section takes the reader through the trading and settlement of futures through a sequence of examples. Not all futures contracts involve delivery of a physical asset in exchange for payment. A certain type of futures contract dictates that delivery be conducted with an exchange of cash. This type of contract is typically referred to as a *cash-settled futures contract*. An example of a futures contract that is cash-settled is a stock index futures contract. Those who are long on a stock index futures contract do not have to accept delivery of the stocks that make-up the index, nor do the shorts have to make delivery. Instead, if the position is held to expiration, the long and short must either pay or receive the difference between the initial entry price and the expiration price. If the futures price increases, then the holder of the long position receives a payment from the short for an amount equal to the difference. If the futures price decreases, the holder of the short position receives a payment from the long for an amount equal to the difference.

Example I.C.6.2: Buying a Futures Contract

Suppose trader A places an order with a futures representative to buy one November flaxseed futures contract on the Winnipeg Commodity Exchange. The order is relayed to the floor of the exchange, where it is filled at a price of C\$420 per tonne. The speculator is now long, and if the contract is held to the expiration in November, he/she is obligated to accept delivery of 20 tonnes of flaxseed from the short based on the terms of the contract at an effective price of C\$420 per tonne.³ The terms of the contract are standardised as to the quantity (20 tonnes) and quality (#1 CW or #2 CW) of the flaxseed that will be delivered and the location(s) to which it will be delivered.

Although a futures contract represents an obligation to deliver or accept delivery of cash or an underlying asset, in 98% of futures trades that obligation is terminated prior to the delivery period through what is known as an *offsetting* trade. Settlement by offset is accomplished by the holder of a long position independently selling the contract, or the holder of a short position independently buying back the contract. The payoff from settling the contract prior to delivery is calculated as the difference between the offsetting and original entry prices.

Example I.C.6.3: Settlement by an Offsetting Transaction

Prior to the start of the delivery month, trader A (who is long one November flaxseed futures contract) places an order with the same futures representative to sell one contract of November flaxseed futures. The order is relayed to the floor of the exchange where it is filled at a price of C\$430 per tonne. By selling November flaxseed, the trader has in effect cancelled out or offset

³ The actual delivery price for a futures contract is the settlement price at delivery. The effective price, however, is the initial entry price as the contract profit or loss is netted from the settlement price.

the earlier long position. As the offsetting price is higher than the original delivery price, the speculator has earned a profit of C\$10 per tonne, which is based on the difference between the buying and selling prices. As a contract represents 20 tonnes, A's profit is C\$200.

Many individuals unfamiliar with the workings of the futures market visualise receiving physical delivery of the underlying asset. Needless to say, the thought of having 20 tonnes of #1 CW flaxseed, for example, dumped on one's doorstep is enough to steer anyone well clear of the futures markets. But, in fact, nothing could be further from the truth. The delivery period only begins with the first delivery day, which is typically near the end of the month prior to the delivery month. As long as a contract is offset prior to this important date (and, as mentioned above, 98% of all contracts are offset) there will be no need for any involvement with delivery. Even if an individual decides to take delivery, what is received/delivered in the case of most physical commodities is a *warehouse receipt* that the seller endorses over to the buyer. The receipt is issued by a storage point, authorised by the exchange, which confirms the presence and ownership of the underlying asset.

Contracts that have not been offset prior to the delivery period are subject to physical delivery (with the exception of cash-settled futures which will be discussed later in this chapter). There are several considerations to keep in mind with regard to physical delivery.

First, it is the short party that controls the delivery process. Within what is allowed by the terms of the futures contract, the short party determines the time and location of delivery as well as the quality or grade of the underlying asset to be delivered. Most contracts allow for multiple delivery points and for the delivery of grades that may be slightly better or worse than what par delivery specifications demand. The allowance of premium or discount grades is designed to increase the amount of a commodity available, and to help prevent one group from controlling or 'cornering' the market.

A second consideration is that the delivery process begins with what is known as *first notice day*. The exact day depends on the particular futures contract, but typically it occurs near the end of the month preceding the delivery month. If a long futures holder such as trader A does not offset a position prior to this day, there is a risk of receiving a delivery notice. The risk grows the further into the delivery month the contract is held. If the contract is held to the end of last trading day, delivery is guaranteed.

At any time on or after first notice day, shorts will notify the exchange's clearing house of their intention to deliver, the location of delivery, and the deliverable grade. Upon this notification, the

clearing house will then allocate delivery notices among clearing members who have long positions on or after first notice day. One method of allocation that clearing houses use is the ‘first in, first out’ method, whereby the oldest long positions are given notices first.

Once the party with the long position receives a notice, actual delivery typically will take place a few days later. On the delivery day, the party with the long position in the contract issues payment by certified cheque to the short position and in exchange takes delivery. Rather than receiving the actual physical commodity at that time, the long will receive a warehouse receipt that represents the amount and the grade of the commodity that is stored at one of the acceptable delivery points. If the underlying asset to be delivered is a financial product such as a currency or bond, in exchange for the certified cheque, the party with the long position will receive documentation that verifies ownership of the asset at an exchange-approved bank. Some futures contracts call for settlement by cash and not by physical delivery. Stock index futures are the most common type of cash-settled futures contract.

Example I.C.6.4: Settlement by Delivery

Instead of offsetting the long position in November flaxseed futures, trader A decides to carry the position past first notice day. In early November, A receives a delivery notice. The notice, which has been delivered to the clearing corporation the previous day by the short, calls for A to accept delivery of a warehouse receipt that represents 20 tonnes of #1 CW flaxseed at an exchange-approved warehouse at a price of C\$440 in two days’ time. On the delivery day, A accepts delivery of the warehouse receipt in exchange for a certified cheque in the amount of C\$8800.

Notice that the delivery process is initiated by the short party, who delivers a notice of intention to deliver to the clearing corporation in early November. In actual fact, the notice will be delivered by the short party’s broker on instructions from the short party. The notice includes details as to the timing of delivery, the grade of flaxseed to be delivered (in this case par value), and the location where the flaxseed is stored. The notice does not specify to whom delivery is to be made. The clearing corporation allocates delivery notices to the various member firms showing long positions. The member firms will then in turn allocate them to their clients who are long.

Readers should note that the delivery price in this example, rather than being trader A’s entry price of C\$420, is actually C\$440, which represents the settlement price on the day the short issued the delivery notice. Based on a price of C\$440, A issues a cheque for C\$8800 to the short (C\$440 × 20 tonnes). While a cheque is issued for C\$8800, the net cost of the flaxseed to A is

only C\$8400. A profit of C\$400 is earned on the long futures position which is automatically closed out the day the delivery notice is issued. The profit is the difference between the settlement price on this day (C\$440) and the entry price (C\$420). The effective net price A pays is C\$420 per tonne (C\$8400/20 tonnes), being the initial entry price.

As has been mentioned, most market participants have no desire to accept or make delivery of an underlying asset. The best way to avoid making delivery is to offset the position before first notice day. If the market participant still wishes to maintain the same exposure to a particular futures contract, 'rolling over' into a more distant contract can be done by offsetting the old contract, while simultaneously entering into a new contract. In the flaxseed example above, if trader A does not wish to take delivery, but wants to maintain a long exposure, the November contract will be sold prior to first notice day while at the same time a deferred flaxseed contract such as March will be bought.

Member firms have procedures for notifying their clients that first notice day is approaching. Typically, the client will be notified several days prior and advised to either liquidate the position or roll over to a more distant month. In order to encourage their clients to offset or roll over their positions, margin requirements are typically raised significantly on and after first notice day. Occasionally, however, a long client, who has no intention of taking delivery, accidentally holds on to the position through first notice day and receives a delivery notice. Most exchanges do have a mechanism that allows those clients to offset their obligation by selling an equivalent number of futures contracts, and then 'passing along' the delivery notice to the clearing corporation, which in turn allocates it to another long position. This procedure, however, can be costly to the client, entailing extra commission costs as well as the possibility of the carrying costs of the physical commodity if the delivery notice cannot be passed on right away.

I.C.6.3.3 Types of Orders

When placing an order in the futures markets, there is some common terminology that is essential to understand in order to be sure that orders are executed properly. Below is a list of some of the most common order types. This terminology applies whether buying or selling a contract.

- *Market order.* This order is used to buy or sell immediately at the 'market price'. There is no guarantee what that price will be, so you rely on the broker and trader for timely and effective execution.
- *Best efforts or worked order.* This order is placed when you wish to give the broker or trader some discretion in executing the transaction. It is often used for large orders where a 'market' order might disrupt trading. Again, you rely on the broker and trader

for timely and effective execution and there is no guarantee as to the price at which the trade is executed, or even if the trade is executed at all.

- *Good 'til cancelled (GTC)*. This is an order to execute a trade that stays 'live' until the customer cancels the trade. Many firms will cancel all GTC trades at the 'close of business', but others will not. It is important to understand the difference in how your brokerage treats GTC trades.
- *Market on open (MOO)*. This is a 'market' order that will be executed when the market opens, at a price within the opening range of prices. Opening price ranges can be quite wide, so this type of trade is to be used with discretion.
- *Market on close (MOC)*. This is a 'market' order that will be executed when the market closes. The price of the trade will be within the closing range of the day, which may be quite large and vary substantially from the settlement price. As with an MOO order, MOC orders are to be used with discretion.
- *Limit order*. This order is placed when you are looking to buy or sell at a specific price 'or better'. This tells your broker or trader in the pit that you are looking to purchase the futures contract at a price no higher than your limit or to sell at a price no lower than your limit. When using this type of order, you should be aware that the market may trade at your limit price for substantial periods of time and you may still not be filled at your order. You are only guaranteed to have your order executed if the market trades through the limit price, either above your sell limit or below your buy limit.
- *Stop order*. This is an order to buy or sell when the market reaches a certain price. Once that price has been reached, the order becomes a 'market order'. A buy stop is placed above the market and a sell stop is placed below the market. Stop orders are commonly used to protect profits or to attempt to limit losses. One should note that markets have a tendency to 'find stops', meaning that when a market price is reached that triggers 'stop orders', the market will often reverse price trends.
- *Market if touched (MIT)*. Much like a stop order, an MIT order becomes a 'market order' if the price reaches a specified level. Unlike the 'stop order', an MIT order to sell is placed above the current market price, and an MIT order to buy is placed below the current market price. Not all firms or exchanges will accept MIT orders.
- *Fill or kill (FOK)*. This order is a limit order that is sent to the pit to be executed immediately and if the order is unable to be filled right away, it is cancelled.
- *Spread order*. A simple spread order involves two positions, one bought and one sold. The trades generally involve the same market with different months (calendar spread) or closely related markets, such as interest rates of different maturities. An order is entered at a 'spread' between the prices of the two contracts. The final execution prices of each contract may not be the same as current 'market' prices of each

individual contract, but each contract's price will be within the day's trading range for that contract.

- *Fast markets.* Of note to anyone who is executing trades on futures markets is a condition known as 'fast markets'. Such a condition means that there is excessive price volatility, usually in combination with a lack of normal liquidity. During 'fast markets', the normal rules that cover whether a trade will be executed according to specific orders are suspended. There are no guarantees of prices or execution. In general, one should be very cautious about entering any type of order during 'fast markets'.

I.C.6.3.4 Margin Requirements and Marking to Market

Futures transactions are typically margin transactions. But unlike margins on securities (which are a counterpart to the maximum loan value that a dealer may extend to its customer to purchase a security), a futures margin is the amount of money that a customer must deposit with a broker to provide a level of assurance that the financial obligations of the futures contract will be met. In effect, futures margins represent a good faith deposit or a performance bond.

The minimum margin rate for a client who wishes to establish a position in a futures market is set by the exchange or clearing house, but a member firm may impose higher margin rates on its clients. The member firm, however, may not charge the client *less* than the exchange's minimum requirements.

Two levels of margin are used in futures trading – *original* and *maintenance margins*. Original or *initial* margin represents the required deposit when a futures contract is entered into. Maintenance margin is the minimum balance for margin required during the life of the contract.

Readers will recall that one of the characteristics of a futures contract is its daily settlement or what is referred to as *marking-to-market*. As mentioned earlier, at the end of each trading day, the long makes a payment to the short or vice versa, depending on the relationship between the current futures price and the initial entry price. In fact, this is a slight simplification. First, the payment is not made directly between the long and short, but takes place between the counterparties' respective investment dealers (member firms) through the clearing house. Second, while the long and short's respective accounts are debited or credited each day by the amount of loss or gain, the party who is in the losing position will only have to deposit additional margin when his or her account balance falls below the maintenance margin level. The margining process is explained in further detail in Section I.C.6.5.4.

I.C.6.3.5 Leverage

Since futures prices only reflect the prices of their underlying interests, the question to ask is why futures trading is considered riskier than trading the underlying interests themselves. The main reason is leverage. Leverage describes the amount of capital that must be put up in order to buy or sell an asset. In mathematical terms, it is simply the ratio of the investment relative to the amount of capital needed to purchase it. If a \$100,000 house is purchased with a \$25,000 down payment and a \$75,000 loan, the purchaser has a 4:1 leverage ratio. Since futures trading requires smaller margins than equity trading, more leverage is available.

Investors can buy or sell equities with margin deposits ranging from 30% to 80%. For example, a \$10,000 long position in a security eligible for reduced margin can be arranged with only a \$3000 deposit. That same \$3000 deposit, however, could secure a futures position with a value of \$100,000 (futures margin requirements are typically from 3% to 10% of a contract's value). If the equity investor sees the value of the stock rise by 10%, the sale of the stock would yield a 33% return on margin. If the futures price increases by the same 10%, the return on margin would be 333%. Of course, leverage would magnify losses if prices moved in the wrong direction.

While leverage is often associated with futures trading, readers should understand that it is not inherent in a futures contract. A futures trader could decide to deposit a contract's full value as margin rather than the minimum margin required. For example, a trader who goes long a gold futures contract could deposit the contract's value of US\$40,000 (100 ounces at an assumed price of \$400 per ounce) as margin. If this decision is made, the trader would not be leveraged at all.

In practice, most traders will take advantage of the leverage that is offered. It is one of the attractions of trading futures. Leverage, however, should be thought of as separate to a futures contract. It is a feature that most participants will exploit, but some may choose not to use. For example, pension funds in Canada are regulated in a way that prevents them from taking leveraged positions in futures contracts.

I.C.6.3.6 Reading a Futures Quotation Page

End-of-day futures quotations are available in most daily financial publications and their websites. Real-time or delayed intraday quotations are available on most exchanges' websites. Table I.C.6.4 duplicates an intraday quotation on live cattle futures from the website of the Chicago Mercantile Exchange (CME); see www.cme.com. The numbers in parentheses indicate that these items are explained in the notes that follow the table.

Table I.C.6.4: Live cattle futures (prices in US cents)

Month	Current Session							Previous Session		
	Open	High	Low	Last	Settle ⁴	Change	Volume ⁵	Settle	Volume	Open ⁶
FEB03	81.350	81.675	81.050	81.675	–	+475	2,936	81.200	3,325	3,328
APR03	77.150	77.375	76.550	77.350	–	+250	10,701	77.100	8,182	53,791
JUN03	70.500	70.900	70.200	70.900	–	+450	3,557	70.450	4,404	22,709
AUG03	67.450	67.800	67.175	67.800	–	+525	880	67.275	779	9,498
OCT03	69.550	69.975	69.525	69.975	–	+350	216	69.625	387	6,337
DEC03	70.850	71.100	70.750	71.100	–	+400	95	70.700	144	3,146
FEB04	72.250	72.300B	72.250	72.250	–	+75	2	72.175	156	1,049

Calculating the value of the underlying interest represented by one futures contract is a relatively simple task. It is just a matter of multiplying the contract size by the latest price. Most of the financial press includes contract sizes within their end-of-day quotations. The exchanges tend to post this information separately from their quotations. In the case of live cattle futures, the standard size of one contract is 40,000 pounds. With the April 2003 contract trading at a price of 77.35 US cents per pound, the value of the underlying interest per contract is \$30,940 ($\$0.7735 \times 40,000$). Open interest and volume figures are analysed quite carefully in conjunction with price movements to give traders an indication of the technical strength or weakness of a particular market.

I.C.6.3.7 Liquidity and Trading Costs

Liquidity, low trading costs and price transparency are some of the main attractions of futures trading. Some of the most actively traded contracts have trading volumes in excess of 200 million contracts per annum⁷ (e.g. Euro-Bund Futures on Eurex and three-month Eurodollar Futures on CME). In such circumstances it is possible to trade large parcels without adversely affecting the price, and bid–offer spreads are minimal. The spread in the liquid contracts is usually the minimum price fluctuation, called a tick. In contrast, some futures contracts are not at all liquid; days may pass when not a single contract is traded. Exchanges often introduce new contracts, of which the majority fail to attract sufficient liquidity. Those that do not attract liquidity are eventually cancelled. Trading costs also can include commissions paid to brokers and exchange/clearing fees.

⁴ The current session's settlement price is not known until the day's trading has concluded. The settlement price should not be confused with the price of the last trade of the day. The exchange's Pit Committee determines the settlement price, which is most often an average of the prices for trades made towards the end of the session.

⁵ The estimated number of contracts that have traded during the current trading session.

⁶ The open interest at the close of the previous trading session. Open interest represents the number of outstanding contracts (i.e., contracts that have not been liquidated by an offsetting transaction or by delivery). Due to the nature of the calculation, open interest is available only after the trading session concludes.

⁷ *Futures Industry*, March–April 2004.

I.C.6.4 Options on Futures

Options on futures contracts were introduced in October 1982 when the CBOT began trading options on Treasury bond futures. These have added a new dimension to futures trading. While both futures and futures options can provide protection against adverse price movements, the purchase of futures options (as with other types of options) provides the ability to both guarantee a purchase or sale price and, at the same time, allow a hedger to participate fully in favourable movements in the price of the underlying asset. Of course, this feature of options comes at a cost – the premium.

Options on futures are just like any other option except that the underlying interest is a futures contract rather than a stock, bond, currency or stock index. A *call option* gives the holder the right to purchase a particular futures contract at a specific price (the exercise price) at any time during the life of the option. A *put option* gives the holder the right to sell a particular futures contract at a specified price at any time during the life of the option. Most options on futures are American style, although some exchanges offer European options on futures (see Chapter I.A.8).

At most futures exchanges the option premium is paid by the buyer at the time of purchase. No margin is required as the losses are limited to the extent of the option premium. The seller of the option must, however, post a margin. Margining arrangements for options will be discussed in Section I.C.6.5.4. There are some exchanges (e.g. the Sydney Futures Exchange) where option purchases, in addition to sales, occur on a margined basis.

Table I.C.6.5 provides an example of contract specifications for FTSE 100 index contracts. Note that the specifications of the options contracts are closely aligned with the underlying futures contract. The exchange (in this case Euronext-LIFFE) has to establish rules for the selection of exercise prices.

Table I.C.6.5: Specifications of FTSE 100 index contracts

	FTSE 100 index futures	FTSE 100 index options (American style)	FTSE 100 index options (European style)
Unit of trading	Contract valued at £10 per index point	same	same
Delivery months	March, June, September, December (nearest four available for trading)	June or December plus additional months such that the nearest three calendar months are available for trading	As for American style
Quotation	Index points (e.g. 6500.0)	same	same
Minimum price movement (tick size & value)	0.5 (£5.00)	same	same
Last trading day	Third Friday in delivery month	Third Friday of the expiry month	As for American style
Delivery day	First business day after the last trading day	na	na
Trading hours	08:00–17:30	08:00–16:30	08:02–16:30
Trading platform	LIFFE CONNECT	Same	Same
Exchange delivery settlement price (EDSP)	EDSP is based on the average values of the FTSE 100 index every 15 seconds between (and including) 10:10 and 10:30 on the last trading day. Of the 81 measured values, the highest 12 and lowest 12 will be discarded.	Same	Same
Daily settlement price	na	The daily settlement price is based on the 16:30 price of the FTSE 100 index	As for American options
Settlement day	na	Settlement day is the first business day after the expiry date	As for American options
Exercise day	na	Exercise by 17:05 on any business day, extended to 18:00 for expiring series on the last trading day	Exercise by 18:00 on the last trading day only
Contract standard	Cash settlement based on the EDSP	Cash settlement based on a daily settlement price for non-expiring series or the EDSP for expiring series	As for futures

Table I.C.6.5 (cont.)

	FTSE 100 index futures	FTSE 100 index options (American style)	FTSE 100 index options (European style)
Exercise price intervals	na	The interval between exercise prices is determined by the time to maturity of a particular expiry month and is either 50 or 100 index points	As for American options
Introduction of new exercise prices	na	Additional exercise prices will be introduced on the business day after the underlying index level has exceeded the second highest, or fallen below the second lowest, available exercise price	As for American options
Option premium	na	Payable by the buyer in full on the business day following a transaction	As for American options

Source: www.liffe.com

Euronext-LIFFE (like many exchanges these days), allows for even greater choice of contract terms through the availability of *flex options*. A flex option is designed to offer the flexibility of the OTC market, but with the advantages that exchange trading brings such as price transparency and reduction in counterparty risk. Participants may request a price quotation on an option with much longer maturity than standard contracts and with the exercise price of their choice.

Table I.C.6.6 provides an example of pricing for FTSE 100 index options as at the close of trading on 16 July 2004. On this day the underlying asset (FTSE 100 index futures expiring on 17 September 2004) closed with a bid–offer spread of 4339–4340. This tight bid–offer spread reflects excellent liquidity, total volume traded for the day being 55,086 contracts. Liquidity is less impressive in the corresponding option contracts, partly because there are so many. Options on futures expiring in August were the most actively traded, with a total of 38,379 contracts. These are split, however, between puts (20,171) and calls (18,208) and 25 strikes. Note that trading is typically the most active for strikes close to the current underlying. Strike prices that are out-of-the-money are more popular than those that are in-the-money (presumably because they are less expensive). The series with the greatest trading activity on 16 July was the call option with a strike of 4525, having total daily volume of 5479. While the bid–offer spread at the close for this series was only 0.5 point, spreads of 3 points or more are common.

Table I.C.6.6: FTSE 100 index option pricing (European-style), August expiry, prices as at 16 July 2004

Underlying asset FTSE 100 index futures expiring 17 September 2004, last trade price 4340													
Settle ¹	OI ²	Total Daily Vol ³	Vol ⁴	Last Trade at	Last Trade	Bid	Offer	AQ ⁵ Bid	AQ Offer	Strike	AQ Bid	AQ Offer	Settle ¹
609.5	-	0	-	-	-	598.5	610.5	3725	0	4.5	-	15:59:06	2054
560.5	-	0	-	-	-	549	561	3775	0	5	-	08:35:34	1532
511.5	-	0	-	-	-	500	512	3825	0	5.5	2.5	-	3069
463	5	0	-	-	-	450.5	462.5	3875	1	7	-	16:27:56	4628
415	-	0	-	-	-	402.5	414.5	3925	2	8	5	15:59:06	22381
367	-	0	-	-	-	354.5	366.5	3975	4	10	5.5	15:50:28	6467
320	22	0	-	-	310.5	315	319	4025	6	12	8	16:07:46	11164
273.5	32	0	-	-	264.5	268	272	4075	9.5	15.5	12	16:29:22	12242
228	30	0	-	-	219	223	213.5	4125	12.5	20	16.5	15:37:13	6103
185	30	0	-	-	175.5	179.5	171	4175	18.5	26	22.5	16:15:03	4469
144.5	479	838	398	13:20:40	142	134.5	138.5	4225	27.5	35	31.5	16:09:16	10527
107.5	150	7	2	15:20:10	101	98	101	4275	40	49	44.5	16:29:22	2461
74.5	3206	525	40	16:22:14	69.5	66.5	69.5	4325	58.5	67.5	62.5	16:22:14	11637
49	1289	297	148	15:40:49	38.5	41.5	44	4375	82.5	91.5	86.5	14:30:25	1639
28.5	5821	1492	5	15:39:16	22	23	26	4425	112	124	118	15:35:38	4074
15	3717	4450	5	16:28:51	12	11.5	13.5	4475	149.5	161.5	156.5	15:35:26	1850
8	11226	5479	5	16:29:00	6	5.5	6	4525	193.5	205.5	199	10:33:53	4075
4	1844	2477	2	15:38:19	2.5	2	4	4575	240	252	-	10:58:15	270
2	6712	218	148	15:40:49	2	0	2	4625	288.5	300.5	-	-	234
1	2554	6	3	08:42:02	1	0	3.5	4675	337.5	349.5	-	-	55
0.5	8793	2373	13	15:28:53	1.5	0.5	3	4725	389.5	401.5	-	10:41:18	74
-	5471	24	10	14:49:42	1	-	-	4775	437	449	-	-	0
-	11053	40	35	14:42:09	1	-	-	4825	486.5	498.5	-	-	0
-	1978	0	-	-	-	0	3	4875	536	548	-	-	0
-	1295	0	-	-	-	0	3	4925	586	598	-	-	5

1. Settle – the previous day’s settlement price.
2. OI (open interest) represents the outstanding long and short positions of the previous trading day updated the morning each day.
3. Total daily volume is the number of trades that have taken place within the respective strike in the trading day. This figure updates as the day progresses and more trades take place within the same strike.
4. Vol (volume) is the number of contracts traded in the most recent transaction.
5. AQ (Autoquote) is the Exchange’s theoretical pricing model for options.

Source: www.liffe.com

When a futures option is exercised, the buyer and writer of the option will receive a futures position in their respective accounts the following day at the exercise price of the option. If a call is exercised, the buyer will receive a long futures position and the writer will receive a short futures position. The entry price for both the long and short position will be the exercise price.

Example I.C.6.5: Exercise of a Futures Option

An investor who feels that canola prices are about to rise decides to buy on 5 November 400 canola call options at a price of C\$2.00. The investor chooses the futures options rather than the outright futures because of the limited risk feature of the former. As with all futures options, the terms of the contract are characterised by the underlying futures. In the case of canola futures, the underlying interest is 20 tonnes of canola. Each dollar move in the price of canola represents C\$20. The total dollar value of the option is therefore C\$40 ($C\2×20) per contract or C\$200 for five contracts. The call option gives the investor the right to buy five contracts of canola futures at the exercise price of C\$400 up to the expiry. If a decision is made to exercise, the call holder will receive five canola futures contracts the day after exercise with an initial entry price of C\$400. If at the time of exercise, the November canola futures price is at C\$410, for example, the call buyer has an immediate open profit of C\$10 per tonne or C\$1000 on five contracts. The buyer can, at that time, decide to liquidate some or all of the contracts and take the profit or maintain some or all of the contracts. If the contracts are maintained, margin has to be deposited. Of course, as with all options, the buyer (or seller) can – and indeed most do – offset the position rather than exercise it. The writer of the call option upon assignment receives five short November canola futures contracts and has an immediate open loss of C\$1000.

Options on futures can be used to either speculate on or hedge an underlying futures contract or the asset underlying the futures contract. An investor who is holding a profitable long gold futures contract, for example, may want to buy a gold futures option put for profit protection (married put). By the same token, an investor who is bullish on gold may just want to buy a futures option rather than buy the outright futures. As far as speculating or hedging cash price movements, the decision to use options on futures or outright futures depends on the investor's risk and return profile. If a limited risk strategy is desired, long futures options would be the choice. If the investor wants to lock in a price with no up-front costs, futures would be the choice.

Example I.C.6.6: Using Bond Futures Call Options to Provide Insurance against Falling Interest Rates

A treasurer of an investment firm that anticipates having funds available at a later date to purchase \$1 million of US Treasury bonds is worried that bond interest rates may decline (bond

prices rise) before they can execute the purchase. The treasurer would like to have temporary insurance against a sudden price increase, but also wants to avoid paying too much if bond prices decline. To achieve these goals, the treasurer can buy call options on Treasury bond futures (each call option represents US\$100,000 par value). Suppose that in May the price of a specific cash Treasury bond is 88–00. Ten September 90–00 futures calls are purchased by the treasurer for US\$10,000. By September, interest rates on long-term Treasury bonds have declined and the price of the cash bond is 96–00, and the ten September 90–00 calls are priced at US\$60,000. The treasurer decides to offset the position by selling the ten September 90–00 calls. The profit on purchase and sale of the calls is US\$50,000 (US\$60,000 – US\$10,000). This profit offsets most of the US\$80,000 increase in the cost of buying the Treasury bonds (US\$960,000 – US\$880,000).

The treasurer in this example, instead of using bond futures options, could have bought bond futures. The futures contract, due to its linear relationship with the underlying asset, would have locked in a price, with no up-front cost, regardless of whether the cash bond moved up or down. If bond futures were used, and bond prices did increase, the futures profit would have offset most if not all of the treasurer's increased cost. The negative side is that the treasurer would not have been able to benefit if bond prices decreased instead of increasing. The savings generated by buying the bond cheaper would have been offset by losses on the futures contract. Of course, the futures contract could have been offset prior to the end of the hedge period, but that would have required the treasurer to go unhedged from that point on.

Here options guaranteed a maximum buying price, but also gave the treasurer the ability to profit if prices declined. The downside of using options is that there is an up-front price to pay. In the example above, it was US\$10,000. Also, an option does not have a linear relationship with its underlying asset the way futures contracts do. If, in Example I.C.6.6, the price of the cash bond rose to 90 at expiration, the option would have no value. The treasurer would not only have paid more for the purchase of the cash bond, but also have incurred a US\$10,000 option loss.

I.C.6.5 Futures Exchanges and Clearing Houses

This section describes the basic features and functions of organised futures exchanges and clearing houses. The trend to electronic trading is one of the most significant affecting the operation of exchanges at the present time. The world's largest futures exchange, Eurex, has *only* electronic trading. With its lower trading costs, it has gained an important competitive advantage, even entering the Chicago market to compete with CBOT.

When an order is placed to buy or sell a futures contract, the order is relayed to the futures exchange where that contract is listed. If the futures exchange uses a trading floor, the order is

relayed to the contract's trading pit. For the more active futures contracts, several hundred traders surround the octagonal-shaped pit. The action around the pit is frantic, with traders shouting, waving their arms and signalling with their fingers orders relayed to them by runners, who run or signal the orders received via phone lines or other communication facilities. Once a trade is consummated between two traders, details are filled out on a trading card and the confirmation is given to the runner who relays it back to the broker who then notifies the client.⁸

If the futures exchange uses an electronic trading system, orders are entered directly into the system where they are matched on a price–time priority, which simulates the auction system used on trading floors. When an order has been filled, the trading system automatically notifies the brokerage firms from where the orders originated. Each brokerage firm would then notify the broker who then notifies the client.

I.C.6.5.1 Exchanges

Futures exchanges provide a forum for market participants to buy and sell futures contracts. Traditionally, this forum consisted exclusively of a trading floor with pits in which traders bought and sold futures contracts through an open outcry auction process. Not so any longer. The late 1990s witnessed a big migration of futures volume from so-called physical trading floors to the electronic trading platforms developed or acquired by most futures exchanges. Some exchanges, like the Bourse de Montreal, went all out and completely shut down their trading floors. Others such as the CME and CBOT created 'side-by-side' trading whereby some of the exchanges' contracts simultaneously trade in both open outcry and electronic trading venues. In these cases, customers are able to direct their orders to either the trading floor or the electronic trading system. A smaller number of exchanges, including the New York Mercantile Exchange (NYMEX), decided to offer electronic trading only in an overnight trading session when the trading floor is closed. A still smaller number of exchanges, the Winnipeg Commodity Exchange among them, have no electronic trading facilities at all.⁹

Regardless of the type of futures exchange, the price buyers and sellers agree upon is arrived at through an *auction process*. The term *open outcry auction process* is used to describe trading on a physical exchange. In this type of auction system, bids and offers are communicated between *floor traders* in a trading ring or pit through both verbal and hand communications. Once a trade is consummated, *market reporters*, who operate from strategic locations around the floor of the

⁸ Modern technology is quickly eliminating, in some cases, the need for runners. Orders are increasingly being entered into trading pits directly.

⁹ In an agreement reached in April 2004 with the CBOT, the Winnipeg Commodity Exchange will be converting all of its futures and options trading facilities from open outcry trading to electronic trading through the facilities of the CBOT's electronic trading platform in the last quarter of 2004.

exchange, record and input the information into a communications system. Once inputted, the price information can be disseminated almost instantaneously around the world.

On an electronic exchange, a specific futures contract's best bid and offer prices are displayed on computer terminals located in member firms' offices. The terminals also allow member firms' traders to enter orders for any contract trading on the system. As orders are entered, the exchanges' trading systems will sort, display and, when the rules of auction trading say so, match them (i.e. create a trade). Only registered members of an exchange have privileges to trade on that exchange. On a physical exchange, there are two types of floor traders. Those who primarily trade on their own account are referred to as *locals*, while those who fill orders from customers are referred to as *floor brokers*. Locals either own an exchange membership, known as a *seat*, or lease one from an owner. The large number of locals that trade on the US exchanges have been among the most significant opponents of the trend away from physical floor trading.

I.C.6.5.2 Futures Exchange Functions

The primary function of a futures exchange is to provide the facilities for the buying and selling of futures contracts through the open outcry auction system. This means providing the physical space, in the case of physical futures exchanges, and the communications infrastructure, for both physical and electronic futures exchanges, to transmit information between the exchange and the rest of the world. In order to ensure the maintenance of fair and competitive markets, exchanges publish and enforce rules and regulations that meet both regulatory and internal requirements.

Another primary function of an exchange is the development of new contracts, and the revamping and sometimes elimination of existing contracts. It is typically the responsibility of the new products committee of an exchange to study the feasibility of new futures contracts in terms of their economic viability. The exchange then submits new contract proposals to the regulatory authority for approval. In addition to recommending new contracts, exchanges are also responsible for establishing the details of all futures contracts traded on its floor. Those details include contract size, delivery standards and location, tradable months, price increments and margin requirements.

I.C.6.5.3 Clearing Houses

Although an exchange provides the setting for the purchase and sale of futures contracts, no money actually changes hands there. Instead, each futures exchange has an associated organisation that takes care of financial settlement, and helps ensure that markets operate efficiently. This organisation, which is called a *clearing house*, can be set up either as a separate corporation or as a department of the exchange. In Canada, the Canadian Derivatives Clearing

Corporation is responsible for clearing Bourse de Montreal futures and option trades. The Winnipeg Commodity Clearing Corporation has sole responsibility for clearing Winnipeg Commodity Exchange trades.

A clearing house guarantees the financial obligations of every contract that it clears. It does this by acting as the buyer for every seller, and the seller for every buyer (*principle of substitution*). A participant who has bought or sold a futures contract has an obligation not to the party on the other side of the transaction, but to the clearing house, just as the clearing house has an obligation to the participant. The existence of the clearing house means that market participants need not be concerned about the honesty or reliability of other trading parties. The integrity of the clearing house is the only issue. As clearing houses have a good record in honouring their obligations, the counterparty risk in futures trading is considered to be negligible. This is one of the principal advantages of futures trading as opposed to OTC trading.

Clearing houses are able to guarantee the financial integrity of futures contracts through a layered system of financial protection. Margin deposits provide the first layer of protection. Parties to a futures trade must deposit an initial or original margin when the contract is first entered. Through the life of the contract, gains are credited and losses debited to the long and short holder accounts on a daily basis. If losses result in an account's net equity (defined as cash deposited plus/minus any open futures positions' profit/loss) falling under the maintenance margin level, the losing party must make a margin deposit to replenish net equity to at *least* the original margin level.

A primary activity of a clearing house is to *match trades* submitted by clearing member firms. Throughout each trading day, clearing members report the details of executed trades, whether they are on behalf of their clients or are on their own accounts, to the clearing house. Once the clearing house verifies the accuracy of all reported transactions, ensures that there is a buy for every sell, and receives original margin from clearing members, it takes over the financial obligations inherent in the futures contract.

The clearing house does not need to know the actual identities of the parties to the transactions. It only needs to know the net positions of the clearing members. The clients are financially responsible to the member firms, while the member firms are financially responsible to the clearing houses. Once a transaction is consummated and confirmed, the clearing house substitutes itself as the buyer for the seller and as the seller for the buyer. This substitution enables the individual trader to liquidate a position without having to wait until the other party to

the original contract decides to liquidate. The trader has in effect bought the contract from or sold it to the clearing house.

It is also the job of the clearing house to ensure that all deliveries are carried out smoothly, as explained earlier in Section I.C.6.3.2. It is important to keep in mind that the principle of substitution does not apply to deliveries. The clearing house merely matches up the buyers and sellers who then can make arrangements for delivery either outside the clearing house or within the clearing house (in which case the clearing house merely acts as custodian). Once the long accepts the delivery notice, the clearing house's obligation is honoured. The clearing house does not take on the obligations of delivery if one side does not satisfy the conditions of delivery. The clearing members must settle any disputes between themselves in accordance with regulatory by-laws. Neither member has any recourse to the clearing house.

I.C.6.5.4 Marking-to-Market and Margin

Suppose that client A has just entered a long December gold futures contract on NYMEX. The contract calls for delivery of 100 ounces of gold. The price that A and the counterparty to the trade (client B) arrived at through the open outcry auction system was US\$385 per ounce. Therefore, A has contracted to buy and B to sell at the maturity of the contract in December 100 ounces of gold at an effective price of US\$385 per ounce. To give each counterparty to the trade a higher level of assurance that the terms of the contract will be honoured, each must deposit margin of, in this case, \$2000 into their respective trading accounts. In turn, the member firm(s) where the accounts are being held will submit the \$2000 to the clearing house.

The \$2000 is *initial* or *original* margin. Futures margins are set at only a small percentage of a contract's underlying value to give market participants, particularly hedgers, reasonable access to a market.

Assume in this example that the *maintenance* margin level is \$1500. The \$2000 that client A initially deposits in the account is equity. If the gold futures price moves higher, net equity will increase. If, for example, the gold futures price rises to \$386 the day after the contract is initiated, A's equity will increase by \$100 ($\1×100 ounces) to \$2100. The \$100 increase will be at the expense of the counterparty to the trade (client B), whose equity will have declined by \$100 to \$1900 (assuming \$2000 was initially deposited). Client A's account will be automatically credited with \$100 by the clearing house and B's debited by the same amount. Client A can withdraw this amount because the account's net equity would exceed the initial margin requirement of \$2000. Client B, however, would *not* have to deposit \$100 into the account because the account's net equity of \$1900 would still be higher than the maintenance margin level of \$1500. Client B would

only have to make a deposit if the account's net equity fell under the maintenance margin level. If, for example, gold futures rise to \$391, B's net equity would fall by US\$600 to US\$1400, which would take it under the maintenance margin level. Client B would then have to make a deposit of \$600 so that net equity is replenished back to the original margin level. The deposit of \$600 would then go from the member firm to the clearing house to A's member firm and finally to A. Member firms and clearing houses typically net out the amount of margin to be paid or received and make this net payment in a lump sum.

It should be noted that if the original and maintenance margin levels were exactly the same, every dollar lost would have to be physically transferred from the losing party to the winning party. This would be quite onerous and difficult to administer. One of the reasons for having a lower maintenance margin level is convenience. It means that clients do not have to run to their respective member firms to make a deposit every time there is a small fluctuation in their accounts.

Daily transference of margin from losers to winners gives the clearing house, in its capacity as third party guarantor and party to the transaction, a high level of confidence that performance will be honoured. Putting clearing houses in an even stronger position to act as guarantor are the *guarantee deposits* which must be maintained by each clearing member. In addition, the clearing house receives income to support its operations by charging fees for clearing trades and for other services performed.

The size of the initial margin will vary according to the contract and the trader's position. Initial margins are determined by the clearing house and may vary from time to time with reference to historical prices and volatilities or in anticipation of forthcoming price-sensitive events. In the case of option contracts, the choice of initial margin is complicated by the fact that option prices are exposed to multiple risks (see Chapter I.A.8), the most significant being changes in the underlying futures price and changes in the volatility. Finally, the appropriate initial margin will depend on the exact position held by the investor. For example, a spread trade (see the discussion in Section I.C.6.7.4) is less risky than a position in a single contract as the two positions partially hedge one another.

To take account of all these factors, many exchanges now use the Standard Portfolio Analysis of Risk (SPAN) framework which was originally developed by the Chicago Mercantile Exchange. The SPAN framework takes a portfolio of futures and options held by an investor and simulates how its value would react to a series of market scenarios. For example, the portfolio might be revalued for 16 different scenarios with various combinations of up/down movements in the

price of the underlying and its volatility. The scenario with the worst outcome for that particular portfolio is used to set the initial margin. Various adjustments may also be made to take account of other factors such as basis risk in spread strategies and extra volatility which is common in the spot contract.

I.C.6.6 Market Participants – Hedgers

The primary function of a futures market is to allow participants who wish to reduce or eliminate risk to do so by shifting the risk to those who want to assume it in return for the possibility of earning a profit. A market participant may need to either reduce the risk of holding a particular asset for future sale or reduce the risk involved in anticipating the purchase of a particular asset. This section covers only the most basic ideas of hedging with futures. More details on hedging with futures can be found in Chapter I.B.3.

A *short hedge* is executed by someone who owns or, in the case of a farmer or miner, anticipates owning an asset in the cash market that will be sold at some point in the future. In order to protect against a decline in price between the present and the time when the asset will be ready for sale, the hedger can take a short position in a futures contract on the same underlying asset which matures approximately at the time of the anticipated sale. By taking this action in the futures market, the hedger will be able to receive an amount equal to the price agreed in the contract, despite the fact that the spot price of the asset at the time of the sale might be considerably different.

Example I.C.6.7 illustrates a short (or selling) hedge example. In this example, a grain elevator locks in its selling price in advance of selling the actual canola. By hedging, the elevator eliminates the risk of reduced profits due to falling prices by the time the physical canola will be sold. As the hedge is lifted at expiration of the futures contracts, the price of the physical canola and the canola futures is the same.

Example I.C.6.7: Short Hedge

On 1 October, a farmer sells 1000 tonnes of canola to an elevator at a price of C\$420 per tonne. The elevator now has an inventory of 1000 tonnes of canola which it expects to sell in December at whatever price is prevailing at that time. In order to protect this inventory against a decline in prices over the next three months, the elevator sells 50 contracts (each contract represents 20 tonnes) of December canola futures at C\$430 per tonne. By December, at the expiration of the futures contract, the price of canola has fallen to C\$400 per tonne, and the elevator sells 1000 tonnes at this price. At the same time as it sells the physical canola, the elevator offsets the short

futures position at C\$400 per tonne. The cash and futures market gain and loss for Example I.C.6.7 are illustrated in Table I.C.6.7.

Table I.C.6.7: A short hedge

<i>Time</i>	<i>Cash market</i>	<i>Futures</i>
October 1	Buys 1000 metric tonnes of canola at C\$420 per tonne	Shorts 50 December canola futures contracts at C\$430 per tonne
December 1	Sells 1000 tonnes of canola at C\$400 per tonne, losing C\$20 per tonne	Offsets the 50 canola futures contracts at C\$400 per tonne
		<i>Profit = C\$30/tonne</i>

Net result: The C\$30 per tonne futures profit more than offsets the C\$20 per tonne cash sale loss. The hedge earns a C\$10 per tonne gross profit. On a net profit basis, the hedge breaks even as carrying costs (assuming the futures are priced at fair value) will be C\$10.

By hedging, the elevator has actually locked in a C\$10 per tonne *gross* profit (i.e. not including the cost of carry) regardless of what happens to prices between 1 October and 1 December, as long as the canola is not sold or the futures contracts offset prior to December. For instance, if canola prices rise instead of falling, the net sale price is still C\$430 per tonne. If canola rises to C\$450 per tonne, for example, the elevator gains C\$30 per tonne on the physical canola sale, but loses C\$20 per tonne on the short futures position. The C\$10 gross profit and net break-even result are locked in regardless of whether prices rise or fall. This is an example of a *perfect* short hedge.

Note that in this hedge the futures contract is offset (on last trading day). The elevator does not deliver canola, but just uses the futures market to lock in a price by taking an opposite position to the canola inventory. In most hedges this is the case. Hedgers use futures contracts, in most cases, not as a delivery mechanism, but rather as a vehicle to offset adverse changes to the price of assets being carried in their normal course of business. As much as possible, hedgers try to fix future sales prices by short hedging, and fix future purchase prices by long hedging.

A long hedge is executed by someone who anticipates buying the underlying asset at some point in the future. In order to protect against rising prices between the present and the time when the asset is needed, the hedger can take a long position in a futures contract on the underlying asset which matures approximately at the time of the anticipated purchase of the asset. By taking this action in the futures market the hedger has fixed the purchase price, even though delivery does not need to be accepted until some point in the future.

The following example illustrates a long hedge where a dental supply company locks in its purchase price in advance of buying the physical silver. By hedging, the company has eliminated the risk of its net purchase price rising between January and April when the silver will be purchased. It can lock in a net purchase price regardless of what happens to prices between January and April, as long as the silver is not bought or the futures contract offset before April.

Example I.C.6.8: Long Hedge

In January a dental supply company estimates that it will need 10,000 troy ounces of silver in April. The firm is concerned that prices will increase in the interim, and would like to lock in a price. The current spot silver price is \$5.00 per ounce. One way of locking in a price that the company has considered is to buy the physical silver immediately and hold it until it is needed. To do so, however, would tie up considerable working capital. Another alternative for the company is to hope that silver prices will fall in the interim. The risk, however, that prices will increase is just too great. The company decides that it is not in the speculation business and chooses to hedge the price risk instead through the futures market. On 25 January, the company buys two April silver futures (5000 troy ounces per contract) that trade on the NYMEX at \$5.20 per troy ounce. Three months later at expiration of the futures contract, the dental supply company buys the physical silver at \$5.40 and offsets the long futures position at US\$5.40 per ounce.¹⁰

The cash and futures gain and loss are illustrated in Table I.C.6.8.

Table I.C.6.8: Long hedge example

<i>Time</i>	<i>Cash position</i>	<i>Futures position</i>
January 25 at	Anticipates the need for 10,000 ounces of silver in April. Current spot price is \$5.00 per ounce	Buys two April silver futures \$5.20 per ounce
April 25	Silver prices rise to \$5.40. The company buys 10,000 ounces at this price	Offsets futures contracts at \$5.40 for profit of \$0.20 per contract

Net result: The company pays \$0.40 more for the silver than expected. It makes a \$0.20 profit on the rise in silver futures from \$5.20 to \$5.40. The effective purchase price is \$5.40 – \$0.20 = \$5.20.

¹⁰ In reality, a delivery notice probably would have been issued to the dental supply company earlier in the month with respect to the long futures positions. For illustrative purposes, we show the long futures position being carried right to expiration day.

In this example, the \$0.40 rise in the price of silver is at least partially offset by the \$0.20 profit on the futures contracts. If instead of rising, silver prices fall, then the net purchase price is still \$5.20. If silver falls to \$4.80, for example, the company pays \$0.20 less per ounce than the January price, but it experiences a \$0.40 loss in the futures contract which has been offset at \$4.80. The net result is the same. By implementing this hedge, the company has locked in a net purchase price of \$5.20.

The hedges demonstrated in Tables I.C.6.7 and I.C.6.8 are examples of *perfect* hedges. They are perfect because the futures price behaves in a way that is expected relative to the cash price. In other words, the basis did exactly as expected: it narrowed to the point where futures prices and cash prices are the same at expiration.

Prior to onset, a hedger will know with certainty that a hedge will be perfect if both of the following two conditions are met:

1. The hedger's holding period matches the expiration date of the futures contract (a *maturity match*).
2. The asset being hedged matches the asset underlying the futures contract (an *asset match*).

Under these conditions, the hedger will know with certainty how the futures contract price will behave relative to the price of the asset being hedged and thereby will have eliminated the risk associated with a future market commitment. On expiration, the spot and futures price will be the same. But if, at the outset of a hedge, at least one of the two conditions above is not met, the hedger is exposed to what is known as *basis risk*, that being the risk of *unexpected* movements in the basis. This does not necessarily mean that a hedge will not be perfect, only that the chances of an imperfect hedge are significant. In retrospect, a hedge may turn out to be perfect even if the two conditions above are not met, if one of the following occurs:

1. A hedge, where the asset being hedged matches the asset underlying the futures contract, is lifted early (the futures contract is either bought or sold), but the basis behaves in a way that was expected by the hedger at the onset of the hedge.
2. The asset being hedged does not match the asset underlying the futures contract, but the basis behaves in a way that was expected by the hedger at the onset of the hedge.

Consider Example I.C.6.7. The three-month canola futures basis is C\$10 per tonne, or C\$3.30 (rounded) per tonne per month. If the hedge is lifted with one month left to expiration of the futures and the basis is at C\$3.30 at that time (as expected), the hedge is still considered to be perfect because it behaves exactly as expected when first implemented. If the canola is sold at

C\$410 at that time and the futures contract is lifted at C\$413.30, the elevator will still break even on a net profit basis. It will lose C\$10 in the cash market, earn C\$16.70 on the futures contract, and pay C\$6.70 (rounded) by carrying the canola for two months.

I.C.6.7 Market Participants – Speculators

Speculators are those market participants who, in the pursuit of profit, are willing to assume the risk that hedgers are seeking to shift. There are several different types of speculator who operate in the futures market. They are distinguished from each other by a number of factors, including the length of time they plan to hold a particular futures position, the amount of profit per position they anticipate, and the amount they are willing to risk.

I.C.6.7.1 Locals

Locals are also referred to as *scalpers*. This type of speculator operates right from the floor of the exchange and has the shortest time horizon of all. Taking advantage of the knowledge and ‘feel’ gained from their proximity to the ‘action’, the local attempts to profit from small price changes that take place in very short periods of time. The time horizon for a local can often be measured in minutes, rather than hours or days. Since the local is only looking to profit from very small price changes, the amount that is typically at risk on any given trade is small. Consequently, a local depends on relatively large volumes to make a successful living and is a unique feature of open outcry markets.

I.C.6.7.2 Day Traders

As the name suggests, day traders are speculators whose time horizon is a single day. Positions taken during a trading day are liquidated by the end of that day. Positions are not carried overnight. Day traders may trade on or off the floor. They are looking to profit from larger price moves than locals, and as a result they are willing to risk more. However, as is evidenced by their desire not to hold any positions overnight, they are not willing to tolerate a lot of risk. Overnight trading can entail considerable risk, particularly in futures contracts whose underlying assets trade 24 hours a day. Foreign currencies, for example, often see their greatest price moves in European or Asian trading. While a speculator sleeps, the value of a particular foreign exchange futures contract can change significantly. By the time North American trading is set to open, the speculator could be greeted with a significant open loss.

Another risk that day traders tend to avoid is that of holding positions going into major reports that could impact the price of a particular futures contract. Day traders involved with grain and oilseed futures, for example, typically go into major supply/demand reports (released on a regular basis by the United States Department of Agriculture) without any positions. These reports are

typically released just before a market opening or just after a market close. Price movements in response to a report have the potential to be very significant, particularly if the data released are different from market expectations.

I.C.6.7.3 Position Traders

This type of trader has a time horizon that can be measured in terms of weeks or even months. Position traders attempt to profit from longer-term price trends. Timing is not as important for a position trader as it is for a local or day trader. The position trader is typically well financed and is therefore in a better position to avoid being *whipsawed* out of the market. This expression refers to a common occurrence in futures trading where a speculator is forced to close out a position due to an adverse price movement, only to see the price quickly rebound back in the favoured direction. Position traders are willing and able to withstand adverse short-term price changes to a larger extent than locals or day traders, in order to maintain a position consistent with their long-term view of the market.

I.C.6.7.4 Spreaders

Spreading involves the purchase of one futures contract against the sale of another which is related in some fashion. Spread traders attempt to identify market situations where the price relationship between two related assets has deviated from its historical norm. When such a situation is identified, the trader will take a spread position designed to profit from a move back towards a level or a range that is more in line with historical performance. The trader does this by simultaneously buying the ‘underpriced’ asset and selling the ‘overpriced’ asset. Spreads can be divided up into four broad categories, as follows:

I.C.6.7.4.1 Intramarket Spreads

Intramarket spreads are also known as *calendar spreads* or *time spreads*. This is a spread which involves the purchase and sale of futures contracts that have the same underlying asset, but different delivery months. They are very popular with agricultural futures where traders speculate on the relative changes in ‘old’ and ‘new’ crop prices.

Example I.C.6.9: Intramarket Spread on Heating Oil

After reading the *Farmers’ Almanac* prediction of a very cold winter, a spread trader implements a spread strategy in November using heating oil futures. The trader feels that strong demand for heating oil during the winter months will force prices of contracts with delivery during this time frame to rise relative to the prices of contracts with spring or summer delivery. The trader buys the February contract at 59 cents per gallon and sells the May contract at 57 cents per gallon. The spread is lifted in January. The February contract is settled at 63 cents and the May contract

settled at 58 cents. The trader earns a profit of 3 cents as the spread widens out from 2 cents premium February to 5 cents premium February. As each cent move represents US\$420 (a heating oil contract is 42,000 gallons), the spread trader's profit is US\$1260.

I.C.6.7.4.2 Intercommodity Spreads

An intercommodity spread is between two different but related futures contracts. The two contracts may trade on the same exchange or on different exchanges. A trader would implement an intercommodity spread when he/she feels that the price of one asset has become under- or overvalued relative to the price of another asset which has a similar usage. For example, both corn and oats, which trade on the CBOT, are used for animal feed. Historically, corn trades at a premium of 50–200 cents to oats. If that spread rises, for example, to 250 cents, a trader may feel there is an opportunity to buy the 'cheap' oats and sell the 'expensive' corn, hoping the spread will move back down to its historical norm.

Perhaps the most popular financial futures intercommodity spread is what is known as the TED spread, which involves the purchase or sale of Treasury bill futures (T) against the opposite position in Eurodollar futures (ED). A trader buys the TED spread by going long on Treasury bills and short on Eurodollars, and he/she shorts the TED spread by going short on Treasury bills and long on Eurodollars. Generally, in times of economic and/or political turmoil, investors seek the safety of Treasury bills which are backed by the US government rather than Eurodollar deposits which are backed only by the bank that issues them. The collective action of investors during these periods forces Treasury bill rates down (prices up) relative to Eurodollars rates. The most extreme widening of the TED spread occurred during the Continental Illinois Bank crisis in 1984, the stock market 'crash' of October 1987, and the Gulf crisis of 1990.

I.C.6.7.4.3 Intermarket Spreads

An intermarket spread involves the purchase and sale of futures contracts that trade on different exchanges, but which have the same underlying asset. Opportunities arise for various reasons. For example, in the case of wheat futures, which trade on the Chicago Board of Trade, Kansas City Board of Trade and Minneapolis Grain Exchange, a spread opportunity may occur because of relative changes in supply and demand conditions of different deliverable grades trading in each respective market.

I.C.6.7.4.4 Commodity Product Spread

This kind of spread involves the purchase or sale of a commodity against the opposite position in the products of that commodity. The most common example of a commodity product spread is the *crush* spread, which involves taking a long position in, for example, soybeans against a short

position in its products, soybean meal and soybean oil. The objective of the spread is to take advantage of any unusual price differences between soybeans, which are not often used in their natural state, and the products they are crushed into – soybean meal, which is primarily used for animal feed, and soybean oil, which is used as a vegetable oil.

I.C.6.8 Market Participants – Managed Futures Investors

Individuals and institutions invest in managed futures products primarily to gain exposure to an asset class that is distinct from the traditional stocks, bonds and cash. Research into managed futures has found that futures are a distinct asset class due to their low correlation with other asset classes. As a result, the addition of futures to a portfolio of other asset classes can provide diversification benefits. Investors looking to diversify their equity and/or bond portfolios are increasingly turning to managed futures products. Barclay Trading Group, which maintains one of the most widely respected managed futures databases, estimates that the amount of assets invested globally in managed products grew from just under \$300 million in 1980 to \$50.7 billion by the end of 2002.

Essentially there are two types of managed futures: *managed accounts* and *managed funds*. Managed futures accounts are used primarily by high net-worth investors and occasionally by institutions. For instance, an investor who wants some exposure to the futures market, but lacks the trading expertise or the time to trade, may give trading authority over to a trading adviser.

Managed futures funds are investment funds that employ strategies using specified derivatives, physical commodities and leverage. Managed futures funds generally focus on a wide variety of market sectors utilising trading styles that may be based on fundamental analysis, technical analysis or a combination of both. Fund managers may base their trading decisions on fully automated computer programs and/or some degree of personal judgement. Fund managers may also trade on the basis of trends, anticipated trend reversals or arbitrage. Unlike equity funds, for example, whose performance may be highly dependent on the direction of the overall stock market, the performance of managed futures is much more dependent on the skills of the manager.

In many ways hedge funds are similar to mutual funds (or unit trusts). Both hedge funds and mutual funds are professionally managed pools of money that may charge investors front- or back-end sales commissions. Both mutual funds and hedge funds charge management fees and can be bought and sold through an investment dealer at a price equal to the funds' net asset value per share. But in contrast to mutual funds, which are generally limited to buying securities or holding cash, hedge funds are structured as limited partnerships that allow the managers to use a

wide variety of alternative strategies and investments. These include derivatives, short selling, leverage, arbitrage, currency trading, and more. Commodity pools are essentially managed futures funds that are structured and sold as mutual funds.

I.C.6.9 Summary and Conclusion

Futures contracts are instruments that allow users to lock in prices on assets that are to be delivered (or cash settled) at some point in the future. When each contract is initiated, there is an entry price which is known as the futures price. Futures contracts are standardised with respect to the delivery date, and the quantity and quality of the underlying asset to be delivered, are traded on an organised exchange, and the transaction is guaranteed by the clearing house of that exchange. A small security deposit is required from the parties to a futures contract known as margin. This margin is held by the clearing house that guarantees the performance of the contract.

Futures contracts are settled mostly by offset and very few involve delivery of the underlying asset. In addition, a unique characteristic of a futures contract is its daily settlement, known as marking-to-market. At the end of each trading day, the long and short position holders' profit or loss is calculated, and the margin account is adjusted accordingly.

Futures contracts cover a wide range of underlying assets such as commodities, stock indexes, interest-rate products and foreign currencies. They are used mainly for hedging purposes. However, since margin requirements are usually small relative to the value of the transaction specified in the contract, they have also become ideal vehicles for speculation.

The relationship between the spot price and the futures price of the underlying asset is determined by what is known as the cost of carry. The cost of carry reflects storage and financing costs with respect to the underlying asset, less income received from the underlying asset during the life of the contract. In the case of assets which are held for other than investment purposes such as commodities, the relationship between the spot and futures prices is also affected by what is known as the convenience yield. The convenience yield measures the benefits from actually holding the underlying asset when unanticipated shortages occur in the marketplace.

Futures markets represent some of the most important forums for financial risk management. Institutional arrangements have been established to minimise counterparty risk and trading costs and to maximise liquidity and price transparency. All these features create an ideal environment for risk management and have contributed to the growing popularity of futures.

The markets for futures and options on futures continue to grow at a rapid pace – volumes have grown in excess of 30% per annum in 2001-03.¹¹ Growth in volumes has been assisted by a number of factors. Countries such as Korea, Brazil, Mexico, China and India have rapidly growing futures markets, and they are now important markets in global terms. Electronic trading, with its low trading costs, has also been a factor supporting the overall market growth. Intense competition between exchanges (e.g. Eurex and CME) has created an environment in which futures markets continue to evolve to respond to market needs and trading costs are kept low. New contracts are regularly launched, and new initiatives introduced to increase market share. An excellent example is the introduction of flex options which allow for greater customisation and therefore reduce the need for OTC trades.

It is likely, therefore, that futures will continue to be an essential tool for the management of financial risk.

¹¹ *Futures Industry*, March–April 2004.

The Structure of Commodities Markets

Colin Lawrence and Alistair Milne¹

I.C.7.1 Introduction

Commodities are traded in both spot and forward markets. They are physical as opposed to financial assets, creating the need for storage and shipping. Because commodities are generally not perishable and can be stored, they are also an asset and can be used as a store of value. Gold and silver have been units of account and numeraires of the entire financial system, as well as a medium of exchange and a store of value. Forward markets for commodities have existed for centuries because, with high volatility, risk-averse producers and consumers have attempted to hedge their inventories in forward and futures markets.

Risk managers in any market have a special interest in the pricing of forward contracts. One key observation of commodities is that the term structure of the forward curve has often been downward sloped, despite the fact that there are non-trivial storage and other transactions costs. This chapter will examine the reasons for this observed *backwardation* of commodity markets. Backwardation is something of a puzzle, since storage costs would normally be expected to raise future prices above spot prices. Explaining backwardation provides an excellent introduction to the special characteristics and risks of commodity markets.

We begin in Section I.C.7.2 by describing the universe of commodities, the various delivery and settlement mechanisms, and the liquidity of their markets. We further examine why gold is special due to its importance as a reserve asset. In Section I.C.7.3 we introduce the reader to the famous arbitrage condition linking forward prices to spot prices and introduce the concepts of convenience yield, backwardation and normal backwardation. Section I.C.7.4 analyses the risks associated with short squeezes and provides some famous illustrations such as the Hunts Silver corner in 1979. Section I.C.7.5 then provides an analysis of downside risk in a typical commodity trading book. Here we show how to estimate the value-at-risk of a commodity portfolio when taking account of the large volatility of borrowing and lending costs. In Section I.C.7.6 we provide some further observations on the behaviour of commodity prices, and Section I.C.7.7 concludes the chapter.

¹ Colin Lawrence is Managing Partner of LA Risk & Financial Ltd, London, and Visiting Professor, Cass Business School, City University, London. Alistair Milne is Senior Lecturer, Faculty of Finance, Cass Business School, London. We are indebted to Carol Alexander and Elizabeth Sheedy for their comments but we remain responsible for all errors.

I.C.7.2 The Commodity Universe and Anatomy of Markets

I.C.7.2.1 Commodity Types and Characteristics²

Commodities are divided into four types: the *metals*, the *softs*, the *grains and oilseeds*, and *livestock*. These generally trade in the spot markets and most have evolved forwards, futures and option-based contracts. The metals can be decomposed into *base metals*, such as non-ferrous metals (e.g., zinc, aluminium, lead and nickel); *strategic metals*, such as bismuth and vanadium, and *minor metals*, such as cobalt and chromium; and *precious metals*, such as gold, silver and palladium. The London Metals Exchange (LME) is one of the key spot-trading centres for both base and precious metals, while strategic and minor metals, having less homogeneity, tend to be traded over the counter (OTC) between producer and consumer. The buyers tend to be automotive, aerospace, pharmaceutical and electrical corporations.

The softs include cocoa, sugar and coffee, and minor softs include rubber, tea and pepper. Most trading of soft commodities involves processors, roasters, refiners, distributors and traders who are 'inventory flow traders' or speculators. The grains and oilseeds category spans most edible agricultural products. It can be further decomposed into the *grains*, such as wheat, barley, rice and oats; *oilseeds*, such as soybeans, rapeseed, palm kernel and flaxseed; *fibres*, such as wool, cloth and silk; and finally, *livestock and other*, including live animals and meat products such as pork bellies. Also included in the latter category are dairy products, such as milk and cheese, and citrus and tropical fruits, such as orange juice.

I.C.7.2.2 The Markets for Trading³

There are two types of markets: the spot commodity market and the market for commodity forwards, futures and other derivatives such as options. Spot transactions take place *on the spot*. They are OTC transactions and could take place in auctions or sales rooms. Once the characteristics are agreed, the commodity is sold and payment takes place at settlement. Usually spot trades involve the exchange of cash for the specified commodity. One must distinguish between the cash for commodity exchange in contrast to settlement of difference. Even a spot transaction takes time to settle, usually in 2 to 45 days. The key problems in the spot or OTC market (which indeed are the key drivers for margin-based structures and exchanges) are: lack of contract transparency, transaction costs, search and time, and creditworthiness. These reasons are the drivers for more standardized contracts and clearing-house management of credit risk in commodities exchanges.

² This section draws heavily on Reuters (2000, pp. 7–129).

³ See Reuters (2000, pp. 31–39).

We also need to distinguish between the markets for forwards and for futures. One important difference between the two is that forwards generally trade in OTC markets. They can be tailor-made contracts, with quantity, quality and maturity all designed to the customer's requirements. Like spot, forward settles cash for the physical. In fact, since a spot trade can settle anywhere between 2 days and 45 days, the spot market is in essence a short-dated forward. Most forwards are not traded on exchanges, but there are anomalies. For example, on the LME, forwards are traded. Futures often (but not always) trade on the basis of *settlement of difference*. This is when the difference, for instance between the forward and actual cash settlement price at maturity, is settled by a net cash payment. Futures are standardised exchange-traded contracts. They have historically been traded in pits under the *open outcry* system. Whilst it looks pretty chaotic, it is a well-regulated and institutionalised form of trading. With the impetus of technology, open outcry trading is giving way to newer forms of electronic trading (see Chapter I.C.1).

Participants in commodity exchanges are brokers and traders acting on behalf of clients, and locals who act on their own behalf. The key distinction between a broker and a trader is that a broker executes on behalf of the client, whereas the trader trades on his own behalf. Often enough, trading and broking are performed by the same firm or individual. There have thus emerged a stringent set of exchange rules which attempt to prevent traders front-running positions or more generally not acting in the best interests of the customer (see, for example, Telser and Higinbotham, 1977). Members of the exchange must execute the orders of customers before they execute their own trades.

The incentive for trading on an exchange is high for all parties. Firstly, trading continually takes place under guaranteed conditions, the price is determined in a transparent manner and, due to posting of margins, counterparty exposure is minimised. The reduction of counterparty exposure through the posting of collateral is the chief function of a clearing house. The key disadvantage is that transactions are standardised and maturity dates are fixed. The standardisation and lower counterparty exposure are critical ingredients in ensuring 'liquidity' of trading, which is the ability to buy or sell a contract with relative low transaction costs.

I.C.7.2.3 Delivery and Settlement Methods⁴

Irrespective of whether the commodity is traded spot, forward or futures, the delivery and settlement methods are critical in determining the actual spot, forward or futures price. There are at least six characteristics of delivery mechanisms:

⁴ See Reuters (2000, pp. 61–129).

1. *In store* is the simplest form of physical delivery. It is used for example, in softs such as coffee and cocoa. *The seller is responsible for delivery to an agreed warehouse.* As in all physical deals, quality, quantity and *location* are all negotiated or embedded in the terms of a standardised (futures) contract. A ‘warrant’ is delivered in the form of a bearer document and is a warehouse claim on a physical commodity. When the trade is concluded the seller transfers ownership of the warrant to the buyer. The product is then shipped to the required location and the buyer exchanges the warrant for the physical or alternatively can transfer it to a third party. Once the contract changes title, the ‘warrant is cancelled’.
2. *Ex store* is identical to in store except that the seller prepays the storekeeper for loading onto the buyers’ transportation. Thus the price will be more expensive in ex store. This process is used in the UK for cocoa and grain. In all physical deliveries, the seller must deliver many documents including a grading certificate, the warrant and weight notes. When these are delivered to the buyer, then the latter pays cash.
3. *Free on board (FOB).* Once the goods have passed over the ship’s rail, the seller has fulfilled his obligation. The onus of risk is shifted onto the buyer once goods are loaded, and hence an FOB price will be cheaper by the insurance premium of damage whilst on board as well as the transportation costs. FOB is used mainly when loading and shipping bulk such as gas oils, sugars and soybeans. Once on board a bill of lading is issued by the seller.
4. *Free alongside ship (FAS)* is similar to FOB. It is a form of delivery where goods are delivered alongside the shipping vessel instead of being loaded. The FAS price will be lower than the FOB price by the cost of loading. In both FOB and FAS it must be specified who pays the tax, import duty, docking fees, value added tax (VAT) etc.
5. *Cost, insurance and freight (CIF).* This involves FOB delivery plus the costs of insurance and transportation. The following simple arbitrage equation relates the FOB to the CIF price:

$$CIF = FOB + F + I,$$

where F is the freight cost and I is the insurance premium. The most expensive element will be the freight cost. Markets have developed in shipping bulk and air cargo. The Baltic exchange is one of the important shipping exchanges. Due to the riskiness of transportation, a futures contract, the BIFEX which trades on LIFFE, has developed, enabling providers of transportation to hedge. The BIFEX is based on the Baltic Freight

Index, which is a composite index of the 11 major dry cargo routes. The contract is settled by difference.

6. *Exchange for futures or physicals (EFP)*. It is possible to swap a physical position for a futures position, and this will be subject to off-exchange negotiation.

I.C.7.2.4 Commodity Market Liquidity

Liquidity, or lack thereof, is of critical importance in the trading of commodities.⁵ We define liquidity as the ability to buy or sell at fair market value without changing transaction costs. A market is liquid when there are lots of sellers and buyers, and where large volumes can be executed with small transaction costs. We would expect bid–offer spreads to be smaller, the greater the liquidity of the market.⁶ In trading forward or futures markets, how liquid is each market?

Table I.C.7.1: Cocoa futures (21 May 2004)

Delivery	Last Trade	Bid	Offer	Trade	Change	High	Low	Settle	Volume	Open	
Month	Time	Price	Price	Price	Price				Today	Interest	Volume/OI
Jul-04	16:50:00	808	810	810	–25	838	795	835	3884	49,370	835
Sep-04	16:50:00	821	824	822	–24	849	807	846	1576	25,114	846
Dec-04	16:49:54	840	843	840	–26	869	827	866	1034	43,174	866
Mar-05	16:49:54	858	861	858	–26	881	845	884	436	25,434	884
May-05	16:09:19	843	885	864	–32	893	861	896	248	9,462	896
Jul-05	15:58:00	856	872	871	–38	875	871	909	50	9,130	909
Sep-05	15:24:59	867	887	900	–19	916	900	919	51	6,464	919
Dec-05	14:59:12	875	898	912	–23	912	912	935	1	6,417	935

Table I.C.7.1 shows a live trading screen of cocoa futures contracts. This gives the contract traded, the time the last trade was executed, the bid and offer prices, the price at which the last trade took place and the price change. The price change is computed by subtracting the last trade price today from the previous day’s last trading price. The previous day’s last trading price is the closing price on 20 May 2004. This is shown under the column labelled ‘settle’ For example, the July 04 contract last traded today at 810, and this is 25 ticks lower than the 20 May closing price of 835. Additionally, the table displays the daily volume and the open interest (the total amount of contracts outstanding), to explore the notion of liquidity.

⁵ For a useful methodology of both theory and practice in measuring liquidity, see Lawrence and Robinson (1995a, 1995b).

⁶ See, for example, Telser and Higinbotham (1977). They examined 23 commodities and found a negative correlation between commissions and volume.

As shown in Table I.C.7.1, the front July contract is the most active with an open interest of close to 50,000 contracts and a current daily volume close to 4000 contracts. The bid–offer spread is at a tight 2 ticks. Actually all the contracts out to March are relatively liquid, with an open interest of over 25,000 contracts. The bid–offer spread increases to only 3 ticks. In general spreads tend to widen out as both volumes and open interest decline. There is very little liquidity from July to December 2005. The bid–offer spread has widened considerably and for the December contract the daily volume is close to zero.

For this reason, hedgers who need to reduce exposure for the second half of 2005 might well trade the March 2005 contract. This involves a trade-off between the benefits of the lower bid–offer spread and basis risk. Thus if a long cocoa producer needs to hedge a sizeable physical position in December 2005, he might in the first instance hedge by shorting the July 04 contract. With this hedge in place he is initially protected from any price collapse until the July expiry. But he has not eliminated his risk, as the basis can shift between July 04 and his delivery date. One strategy would be to attempt to roll over his position into the long-dated contracts. He would execute this by buying the July 04 to Dec 05 spread. This is tantamount to closing out his July 04 position and shorting the Dec 05s. He could therefore buy the July 04–Dec 05 spread, which would leave him with a perfect hedge. But transactions costs are extremely high. The bid–offer spread on Dec 05s is 14 ticks (roughly 7 times as large as the front contract). So he is more likely to put in a spread buy order for, say 70 ticks, and execute this quite slowly, hoping there is a ready seller of the spread. The speed at which he rolls over his trade will depend on the trade-off between transaction costs and volatility of the basis. The more volatile the basis, relative to the bid–offer spreads, the faster will be the rollover the position.

In Table I.C.7.2 we depict the open interest and volumes in gold and silver. Just like cocoa, there are large volume trades in the gold front June 2004 contract, but silver’s major liquidity is in the December contract. As we move out the term structure of these metals, we note that the volumes and open interest decline precipitously. However, while volumes decline on contracts further out, there is still significant open interest that grows more slowly.⁷

⁷ Telser and Higinbotham (1977) find that as volumes and open interest increase, commissions decline. Furthermore, that as volatility increases, the ratio of volume to open interest also declines. The key driver for this is that uncertainty induces heavier turnover of positions.

Table I.C.7.2: Liquidity in gold and silver futures contracts (13 May 2004)

Gold			Silver		
Date	Open Interest	Volume	Date	Open Interest	Volume
Jun-04	155,974	53,949	May-04	472	96
Aug-04	22,099	4,837	Jun-04	16	0
Oct-04	6,902	1,484	Jul-04	61,067	13,423
Dec-04	30,070	1,330	Sep-04	5,727	278
Feb-05	3,065	109	Dec-04	14,327	465
Apr-05	3,596	132	Jan-05	6	0
Jun-05	13,421	205	Mar-05	3,716	46
Aug-05	1,888	25	May-05	908	45
Oct-05	276	0	Jul-05	1,201	0
Dec-05	6,511	10	Sep-05	85	0
Feb-06	413	0	Dec-05	2,946	12
Jun-06	6,325	0	Jan-06	0	0
Dec-06	1,815	0	Mar-06	0	0
Jun-07	1,141	0	Jul-06	31	4
Dec-07	1,246	0	Dec-06	1,213	1
			Jul-07	58	1
			Dec-07	63	0
TOTAL	254,742	62,081		91,836	14,371

It is imperative that risk managers incorporate the bid–offer spreads and illiquidity of longer-dated contracts into measures of value-at-risk and stress testing.⁸ And as we shall see in Section I.C.7.4, there is a further very serious liquidity risk which is hidden from these charts, and that is the squeeze or cornering risk in which the shorts have to deliver physicals as the front contract expires. Thus there is an implicit demand for warrants of the physical by the shorts who have to cover their positions. What happens when these inventories become scarce? This has happened throughout history as in the case of silver in 1979–80, when the silver price skyrocketed to US\$50 per ounce due to a squeeze. This type of backwardation happened throughout the nineteenth and twentieth centuries, and Section I.C.7.4 examines these volatile price jumps and the ensuing limits imposed by regulators and exchanges.

I.C.7.2.5 The Special Case of Gold as a Reserve Asset ⁹

Gold has special characteristics relative to other commodities because it is still a reserve asset of central banks and has a history of being pivotal to the international monetary system. As a

⁸ See Lawrence and Robinson (1996) for a comprehensive methodology for working bid–offer spreads and illiquidity into value-at-risk measurements.

⁹ This section draws heavily on “Gold as a Reserve Asset”, World Gold Council. See www.gold.org.

consequence of the world's central banks' holdings of gold reserves, changes in central banking portfolio behaviour can have profound effects on the gold price.

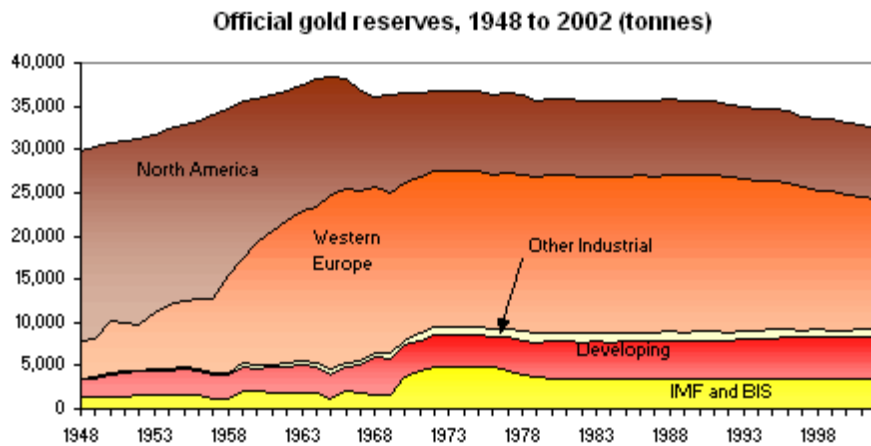
The international gold standard existed for a short time period from 1870 until the beginning of World War I.¹⁰ This period was a relatively stable period, coupled with strong economic growth. Money was fully backed by gold, and investors were protected from inflation. It also meant that international investment was safer than under floating exchange rates. Capital flows were excessively large during the imposition of this standard. A key reason for this was the credibility of the regime. The problem with the standard, however, was that much depended on getting supplies out the ground, and indeed the demand for money could well have outstripped supply at times. The lack of credibility was widespread after World War I. Despite an initial fixing of US\$20.67 per troy ounce, this was suspended in 1933. This culminated in one of the greatest hyperinflations, especially in the Austro-Hungarian Empire. As long as governments used money to finance budget deficits, inflation was a direct outcome. In 1934 gold was fixed at \$35 per troy ounce.

In 1944 the Bretton Woods System of fixed exchange rates was created, with the US dollar as the fulcrum of the system. The dollar was pegged at \$35 dollars per ounce, member nations of the International Monetary Fund had to deposit a share of gross domestic product in gold, and the major central banks all used gold as the core reserve asset. But the system came to a halt following the Vietnam War when the US government under Nixon abandoned the gold standard completely, and since 1971 the price of gold has floated freely. Figure I.C.7.1 depicts the inventories of gold reserves held by central banks and supranational institutions. The USQ has dramatically reduced its reserves from over 20,000 tonnes in 1950 to around 9000 tonnes in 1971, while Europe built up its reserves. At the end of 2002, the USA held about 8000 tonnes and Europe about 13,000 tonnes.

The total inventory of central banks' holdings is around 32,000 tonnes, about 22% of the above-ground reserves. This is huge relative to annual production, which is estimated at around 2600 tonnes. As a consequence of this, speculators and hedgers of gold scrutinise the behaviour of the central banks. On 26 September 1999 the Washington Agreement on Gold was announced, whereby it was agreed that uncoordinated selling of gold by central banks would destabilise its price. The Accord reached was a pledge not to sell more than 2000 tonnes as a 'collective' over the next five years. The accord halted the decline in the price of gold at around \$250 per ounce and certainly contributed to its rise to \$400 per ounce.

¹⁰ See www.gold.org for a chronology of events concerning gold as a reserve asset.

Figure I.C.7.1: Official gold reserves, 1948 to 2002 (tonnes)



Source: World Gold Council, www.gold.org

Despite the fact that gold standard is a historical event, movements in inventories of central banks are critical not only in changing supply in the market but also in determining lease rates (the rates at which banks lend out their gold) and hence the gold basis. Central bankers lend gold to gold dealers with mismatched books in order to earn a return on their gold holdings. There is a ready demand for gold loans (from gold producers) as a cost-effective means of financing their gold production. The gold interest rate paid on gold deposits is called the gold lease rate or gold LIBOR. Thus while most commodities incur storage costs, gold deposits earn a positive return.

To summarise, gold is unique in the commodity universe for the simple reason that the above-ground reserves are so massive and they can earn a positive return. The large scale of these reserves makes it extremely unlikely that a short squeeze (caused by excessive short-term demand) could ever occur. This has enormous implications for the spot–forward pricing relationship which is the subject of the next section.

I.C.7.3 Spot–Forward Pricing Relationships

We have described cash, forward and futures markets for commodities. In this section we review the relationship between futures/forwards and spot prices, sometimes referred to as ‘basis’. Unlike ‘normal markets’ where forward prices are generally higher than spot, forward commodity prices are often found to be below the spot price. Not only does the basis for commodities differ from most other markets, but it is also highly variable over time because of the possibility of short squeezes. This feature of commodity markets is a concern for risk managers because it can make hedging strategies less effective and create unexpected changes in the value of a portfolio (possibly giving rise to significant margin calls). Thus non-normal distributions, stress

testing, and scenario analysis prove to be critical ingredients in the commodity risk manager's toolkit.¹¹

We will ignore the institutional differences and treat the forwards and futures markets as equivalent.¹²

I.C.7.3.1 Backwardation and Contango

The relationship between forward and spot prices has been developed in Chapters I.A.7 and I.B.3. Based on the analysis of typical markets, one might expect the pricing relationship for commodities to be defined as

$$\text{Forward price} = \text{spot price} + \text{carrying cost}, \quad (\text{I.C.7.1})$$

where the carrying costs involve storage costs, transaction costs and all other characteristics in Section I.C.7.2.3, especially freight and interest costs.

In studying patterns of spot versus forwards, or short-dated versus longer-dated forwards, we find by empirical observation that the spot price often lies *above* the forward price. See, for example, the sugar and wheat prices in Table I.C.7.3. Indeed in many commodities we find that the forward can switch from being at a premium to a discount and vice versa. If the forward or future is traded at a price higher than the spot price, then we say that the forward or future is at a *premium* or (equivalently) that the market is in *contango*. If the forward is below the spot price, then we say that the market is in *backwardation*.

Table I.C.7.3 shows closing prices of four commodity futures: robusta coffee, white sugar, feed wheat and milling wheat. Robusta coffee is in contango, while both white sugar futures and feed wheat futures are in backwardation; milling wheat shifts from contango to backwardation in July. The daily percentage change is also shown for three of the products. A key observation is that the changes are almost invariably non-uniform across the term structure. For example, March coffee has fallen by 1.71%, while July coffee has fallen by only 0.67%. The reason for the non-parallel shift is the relationship described in equation (I.C.7.1), namely that shifts in forward prices are due to both the change in the spot price plus a shift in carrying costs. The non-parallel shift must be caused by changes in carrying costs.

¹¹ See Section I.C.7.4 for a historical overview of short squeezes, in particular the famous silver squeeze of 1979–1980 manipulated by the Hunt brothers.

¹² The reader is referred to Chapter I.A.7 for an analysis of the critical differences between the two types of markets. General historical and cross-sectional data can be readily obtained from the commodity futures exchanges.

Table I.C.7.3: Closing futures prices¹³

Robusta Coffee			White Sugar			Feed Wheat			Milling Wheat		
Future	Price	Change	Future	Price	Change	Future	Price	Change	Future	Price	Change
Mar-04	680	-1.71%	May-04	206.8	0.63%	Mar-04	95		Mar-04	148.5	
May-04	723	-0.82%	Aug-04	200.5	0.65%	May-04	95.25	-0.42%	May-04	151	
Jul-04	741	-0.67%	Oct-04	194.2	0.67%	Jul-04	97.5	-0.77%	Jul-04	135	
Sep-04	758	-0.91%	Dec-04	194.2	0.83%	Sep-04	77.5		Sep-04	122	
Nov-04	774	-0.89%	Mar-05	192.2	0.36%	Nov-04	78	-0.64%	Nov-04	124	
Jan-05	789	-0.63%	May-05	194.2		Jan-05	–		Jan-05	124.5	
Mar-05	802	-1.35%	Aug-05	192		Mar-05	–		Mar-05		
Term Structure	Contango		Backwardation			Backwardation			Mixed		

The data in Table I.C.7.3 show how the basis can vary from day to day. A key point driving the yield curve or basis is that commodity buffer supplies can run out. We cannot import spot commodities from the future. The risk manager will have to scrutinise demand and supply data carefully. In the case of gold, we noted in Section I.C.7.2.5 that above-ground stocks are very large relative to flows. Hence any incipient excess demand can be met through an elastic supply of inventory, thus mitigating the kind of backwardation we observe in cyclical commodities such as oil and aluminium.

Gold is an exceptional commodity for the reasons cited in Section I.C.7.2.5. The gold market is almost always in contango, as shown in Section I.C.7.3.3. But with a lack of inventory to cope with unexpected changes in the flow demand or supply, other commodities can experience shortages – and hence backwardation. For example, in the last decade there has been a surge in food consumption in China and, with her low area of green land per capita, imports fill the deficit. In the 1990s analysts were watching the near monopolisation of grain containers out of Chicago by the Chinese. Chinese demand pushed prices of food products up by over 13% in 1997. Any gyrations in demand for imports can tilt the yield curve of foodstuffs.

Finally, backwardation shifts are caused by combinations of scarcity and further manipulation of the market. To illustrate the latter point, the LME were recently concerned when the aluminium basis backwardation flared from 2% to about 5% in 2003.¹⁴ On investigation, they found that the market had been in over-supply since 2000. There was a surplus of 1 million tonnes in 2002 and about 400,000 tonnes in 2003; inventories had risen to over 3.5 million tonnes in July 2003, and a large share of this was in LME warehouses. There is a strong possibility of market collusion still being investigated by the LME.

¹³ The prices selected are the bid prices for robusta coffee, white sugar and milling wheat and offer prices for feed wheat.

I.C.7.3.2 Reasons for Backwardation

Keynes (1930) explained backwardation by the presence of speculators transacting in commodities in the hope of making short-term trading profits. These speculators should be viewed as selling insurance to hedgers. According to Keynes, speculators would demand a premium for providing this insurance, and this would generally drive the spot price above the forward price. This hypothesis has undergone a barrage of scrutiny by economists. At first glance it is puzzling how spot prices lie above forwards unless carrying costs are somehow negative. Indeed, for over a century economists have debated why the forward price could lie below the spot price.

Modern portfolio theory and subsequent empirical testing have refuted the Keynesian view of normal backwardation. Efficient market hypotheses have also rejected the premise of Hicks (1946) that forwards are biased estimators of expected spot prices. In fact, the theory which most successfully explains backwardation is the *convenience yield theory* (Kaldor, 1939; Williams, 1986). This focuses on the physical demand and supply of commodities as an input and/or output in the production process. It suggests that spot prices are driven above forwards because users of commodities cannot afford to run out of inventory. They are prepared to pay more in spot markets than the expected future price, in order to ensure that they have immediate access to supply (this is the *demand for immediacy*).¹⁵ From the perspective of suppliers, they are also prepared to hold more than required inventory in case the market switches from backwardation into contango. Despite the apparent negative return to suppliers, there is a chance that they can get exceptionally high returns from shortages, where producers will borrow commodities and thus drive spot prices very high.

Furthermore, a major cause of sudden switches from contango to backwardation has been the fact that the deliverable commodity to any futures contract at maturity can be subject to manipulation. The examples that stand out include the infamous silver corner in 1979–80, the copper scandal in 1996 and the aluminium backwardation in 2003. Whilst the exchanges and regulators have attempted to stop these ‘corners’ with an array of limits on movements in prices per day as well as ensuring the maximum open interest held by one investor is limited, history has been scarred by the omnipotence of squeezes. In Section I.C.7.4 we explore these squeezes in more detail. They can be an Achilles heel for the risk manager and often they are very difficult to detect, and yet they could wipe out economic capital in one stroke!

¹⁴ The spot price of aluminium shot up to US\$70!

¹⁵ Williams (1986) gives a persuasive and cogent analysis of the demand for immediacy and backwardation. Lawrence (2003) shows, however, that there is negligible backwardation in commodities such as gold where inventory supplies (above-ground stocks) are very large relative to production (flows), regardless of the convenience yield.

Convenience yield theory suggests that shortages of physical delivery arise through either an unexpected shock or an anticipated (or unanticipated) act of Nature such as the seasonal harvest. Producers are willing to hoard commodities since they could otherwise suffer an opportunity loss if they do not have the commodity, especially if it is a key input in a supply chain. A consequence is that there are always premium prices at which producers will be willing to hold excess inventory even if they incur all the carrying costs.¹⁶ In such circumstances they could earn huge returns if there is a market shortage and this creates a backwardation.¹⁷

Empirically, commodity markets can switch from contango to backwardation and back to contango, creating much volatility. Thus it is crucial for the risk manager to understand exactly what factors cause such volatility in prices. In Section I.C.7.4 we give some current and historical examples of short squeezes to be a major culprit of backwardation. Indeed, in stress situations real and artificial physical shortages are a major cause of backwardation.

I.C.7.3.3 The No-Arbitrage Condition

In an efficient market with no arbitrage, forward term structures such as those shown in Tables I.C.7.1 and I.C.7.3 will be influenced by the interest rate, storage costs and the convenience yield. To see this, first note that the key arbitrage equation for commodity pricing is:¹⁸

$$(E(S) - S) / S = r + C - q, \quad (\text{I.C.7.2})$$

where S is the spot price, $E(S)$ is the expected spot price, r is the risk-free rate of return, C is the storage cost (quoted as a percentage of the spot price) and q is the convenience yield. This is similar to the arbitrage equation for equities where the storage cost is zero and the convenience yield is analogous to the dividend yield (see Chapter I.A.8). The left-hand side of equation (I.C.7.2) is the expected capital gain, while the right-hand side is the *carrying cost*. The higher the convenience yield, the lower will be the expected capital gain. If $q > r + C$ then the carrying cost can be negative.

If the expected capital gain from holding inventory is greater than the carrying cost then investors or producers will hoard inventory. Then, even if the market price is expected to fall, investors may hold stocks if the convenience yield is high enough to compensate for the loss.

Note that if markets are efficient, $E(S)$ is equal to the future price F . Then (I.C.7.2) becomes

¹⁶ See Telser (1958) for a derivation of a commodity model with speculators and the behaviour of storage.

¹⁷ Normal backwardation, according to Keynes (1930), implies that the *forward price will lie below the expected spot price*. This hypothesis has been thoroughly scrutinised by economists. In the next section we will describe the key arbitrage equation linking spot and forwards/futures and how backwardation and normal backwardation come into play.

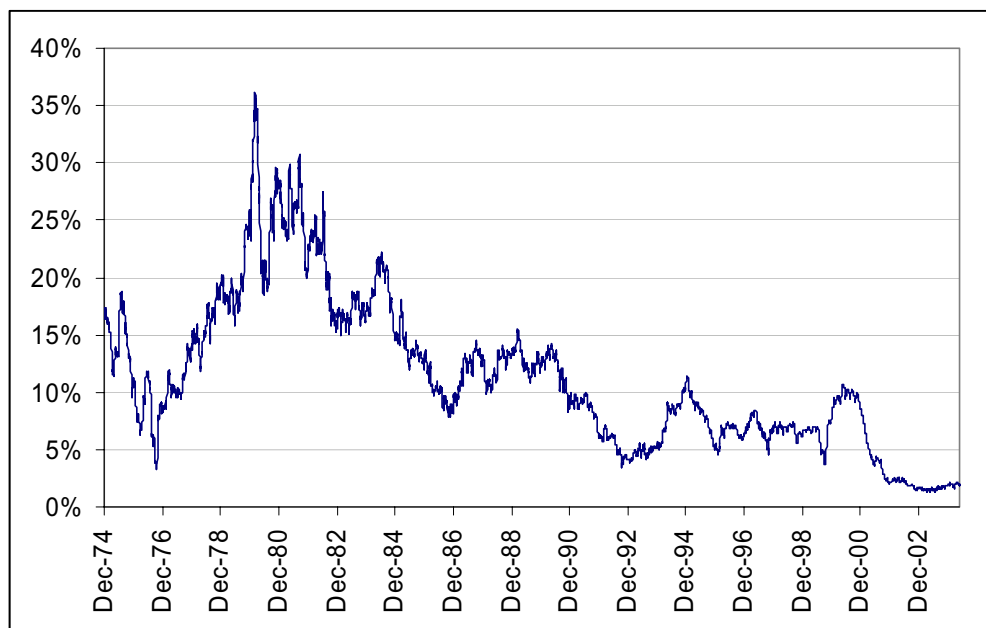
¹⁸ In this equation Kaldor and others have included a risk premium. We have omitted it since we believe that the evidence tends to refute it and thus is not that critical to incorporate it.

$$(F - S) / S = r + C - q. \quad (\text{I.C.7.3})$$

Hence, the future price will lie below the spot price – and the market will be in backwardation – if the convenience yield q outweighs the other two carrying costs. The left-hand side of equation (I.C.7.3) is the *basis* (the spread between the future and the spot price) as a percentage of the spot price. It increases with r and C and decreases with q . More importantly, q is much more volatile than either r or C and so the changes in the convenience yield dominate the changes in the basis. This is a very useful insight. Variations in the carrying cost of commodities can be estimated using variations in the basis in the forward market.

The lease rate of any commodity is depicted by the right-hand side of (I.C.7.3). It is quite difficult to find data on storage costs and difficult econometrically to estimate the convenience yield. However, under the assumption that arbitrage condition holds, then we can proxy the gold lease rate via the futures contango depicted on the left-hand side of (I.C.7.3). Futures and cash data are readily available. In Figure I.C.7.2 we show the lease rate of gold over the period from 1974 to 2003.¹⁹ It is clear that the market is rarely in backwardation – indeed, the graph shows that from 1974 to 2003, gold was in contango throughout. When gold roared in the 1980s so did the contango, reflecting inflationary expectations and a bubble. Lease rates rose to over 30%. But as inflation subsided after the Fed’s tightening policy of 1979, lease rate collapsed to around 1 to 1.5% in present times.

Figure I.C.7.2: Six-month gold lease rate, 1974–2003 (quoted as an annualised yield)



I.C.7.4 Short Squeezes, Corners and Regulation

I.C.7.4.1 Historical Experience

Commodity markets throughout history have been plagued by ‘short squeezes’. The reason for a squeeze in the futures or forward markets is that the delivery of the physical to ‘longs’ by the ‘shorts’ is in an artificially short supply. We now examine some of these historical examples, focusing on the following questions:

- a) Was there a fundamental exogenous shock such as a poor harvest or under-supply or increase in demand that drove prices up and created a backwardation situation?
- b) Did any institution or individual or associated²⁰ party hold a long futures position and attempt to buy spot commodities prior to expiration?
- c) How were these inventories financed?
- d) What was the response of the exchange or/and the regulators?
- e) What was the path of the price before and after expiration?

Clearly, the astute commodity risk manager should ask all five questions. Failing to do so can lead to substantial unanticipated risk exposure.

Example I.C.7.1: Benjamin Hutchingson and the wheat squeeze, August 1866

It all began with poor harvest forecasts. Hutchingson built up long wheat positions in cash and futures markets. The average price was around 88 cents per bushel. In August the spot price started rising in Iowa and Chicago and by 4 August the price had hit 92c a bushel. But on expiration, Hutchingson’s right to demand physical delivery raised the price to \$1.87 per bushel. The Chicago Board of Trade (CBOT) deemed such transactions as fraudulent and declared that any member of the CBOT engaged in this activity would be expelled. This did very little to prevent transactions of these types (see Sikorzewski, 2001, pp. 2–3).

Example I.C.7.2: The Great Chicago Fire and John Lyon, 1872

On 6 October 1872 six out of seventeen elevators burnt down in the Great Fire of Chicago. The storage capacity of Chicago fell from 8 million to 5.5 million bushels. An important merchant named John Lyon formed a coalition with two other brokers. In spring 1872 they began buying up physicals and futures. In July the August contract rose to \$1.16 and by month’s end to \$1.35. The price stimulated a huge inflow of wheat into Chicago. Initially the train volume was around 14,000 bushels, and it rose steadily to 27,000 bushels in the first week of August. A further

¹⁹ Note that one would proxy r by the interbank dollar rate such as LIBOR. The difference between the interbank rate and the LIBOR is measured by $q - C$. Since the volatility of C will generally be small, the difference is a good proxy for the convenience yield.

²⁰ Associated in this context could be a family member, a subsidiary, an insider collusive customer, or any entity that is ‘party’ to the squeeze.

disaster struck when the Iowa elevator also burnt down, reducing storage by 300,000 bushels. Meanwhile further bad weather reports led to rumours that the new crop would mature too late for delivery into the August futures contract. The August futures price climbed to \$1.51 on 15 August.

As a consequence of the price hike, farmers started shipping wheat to Chicago at an accelerating rate. In the first week in August about 75,000 bushels per train reached to Chicago, and by the 19th daily arrivals reached up to 200,000 bushels. The wheat coming in from Buffalo would usually have begun depressing the price. At this point, Lyons and his partners, who had been borrowing heavily from the banks to execute his cash trades and post margin, were turned down for more financing. Meanwhile, the building of new elevators raised the capacity to 10 million bushels, which was higher than the pre-fire capacity. When the banks turned him down on 19 August, the price plunged to 25c per bushel. Lyons then announced that he was bankrupt and the price plunged a further 17c. This and similar squeezes – there were 15 such cases – ultimately led in 1922 to Congress passing legislation in the form of the Commodity Exchange Act outlawing squeezes.

Example I.C.7.3: The great Silver Squeeze by the Hunt brothers, 1979–80²¹

In the summer of 1979 the Hunt brothers, together with their Saudi partners, began buying up silver. They bought about 43 million ounces of silver on the COMEX with delivery to be taken. Over the summer the silver price jumped from \$8 to \$16 as a result of their actions. Other syndicates began to copy them. The COMEX and CBOT were in a panic. In 1979 the warehouses of the two exchanges only held 120 million ounces and that amount was traded in October alone. The Hunts moved a further 9 million ounces of silver to Europe through a silver swap. They were worried that US government would attempt to confiscate their inventory. Late in 1979 the CBOT changed its rules and stated that no investor could hold more than 3 million ounces of silver contracts and any investor with over 3 million would have to liquidate by February 1980. The margin requirement was also raised. The Hunts accused the exchange of vested interests. This was borne out by the fact that many CBOT members had ‘shorts’. The Hunts, knowing that indeed the market was synthetically short, continued buying silver. At year’s end the price had risen to \$34.45 per ounce. The brothers at this time held 40 million ounces in Switzerland and 90 million ounces through their holding company, International Metals. On top of that they had longs for the March contract on another 90 million ounces. Lamar Hunt, the younger brother, had also accumulated an individual position \$300 million position by the end of 1979.

²¹ For a delightful analysis of all the corners, scandals and squeezes in historical and modern times, see Geisst (2002). For an exposition of the Hunt silver scandal, see pp. 212–241.

Finally, on 7 January, the COMEX changed their rules to only allow 10 million ounces of contracts per trader and to require all traders to reach this threshold before 18 February. The Commodity Futures Trading Commission (CFTC) backed the ruling.²² On 17 January silver had peaked at \$50. Amazingly, the Hunts continued buying – their position was worth \$4.5 billion and their profits were estimated at \$3.5 billion. On 21 January, the COMEX suspended trading in silver. They would only accept liquidation orders. Silver dropped to \$39 per ounce and stayed there until the end of January. About 22 million ounces of scraps, including silver coins, silverware and jewellery, came to market. In early February the Hunts took delivery of another 26 million ounces from Chicago. The Hunts had substantial resources – their oil company, Placid Oil, was generating about \$200 million in profits from North Sea oil. There were talks of taking over Texaco and rumours that their Middle Eastern partners were putting together a syndicate to buy silver. A new shock occurred. Paul Volker, the Fed chairman, had decided that inflation had gone too far and pushed up the Fed funds rate to over 21% as he tightened monetary growth. By March silver plummeted as the dollar roared, reaching \$24 per ounce by 14 March. Margin calls cost the Hunts about \$10 million daily. They scrambled across Europe searching for buyers, but as the price dropped not only did their margin calls rise, but so did the cost of margins as a consequence of the higher interest rates.²³ It was all over on 25 March 1980 as they defaulted on a margin call to the tune of \$135 million. By 27 March the price fell to \$10.80 an ounce. They lost \$1.35 billion and the Fed, worried about systemic risk, was willing to extend loans to the beleaguered brothers.²⁴

I.C.7.4.2 The Exchange Limits²⁵

The exchange rules, prodded by the CFTC, have changed dramatically. All commodity contracts have two sets of limits: the first are daily maximum movements in price of any contract; and the second are limits on each individual trader. Column 3 in Table I.C.7.4 shows the daily limits on a sample of commodities traded on a number of exchanges. For pork bellies, for example, the contract size is 40,000 lbs. The minimum daily move is US\$4.00 per contract and the exchange will suspend trading if the contract falls or rises by 200 ticks.

²² The CFTC is the key regulator of commodities, exchanges and derivatives in the USA. Its key mission is to protect market users and the public from fraud, manipulation, and abusive practices related to the sale of commodity and financial futures and options, and to foster open, competitive, and financially sound futures and option markets (see <http://www.cftc.gov>).

²³ The risk manager will be aware that this is a ‘negative convexity’ stress exposure. In this case both the cost and the quantity of margin demanded have risen at heightened rates.

²⁴ Years later in 1998 the precedent set by Volker of lending the Hunts money to avoid systemic risk was emulated by Greenspan arranging support for LTCM. (See United States Treasury report on Hedge Funds, Leverage and Lessons from Long Term Capital Management, Pages 1-42. www.treas.gov/press/releases/reports/hedfund.pdf and Shirreff on Lessons from the Collapse of Hedge Fund, Long-Term Capital Management. <http://newrisk.ifci.ch/146480.htm>)

²⁵ For an insight into the setting of limits and the relationship between the exchanges and the CFTC, see Geisst (2002).

In addition to the limits shown in Table I.C.7.4 there are a range of other restrictions. The most important control is the maximum exposure per trader. In pork bellies, for example, no trader is allowed 1000 contracts net long or short in all contract months, and no trader can have (or control) over 800 contracts long or short in any one month. Furthermore, to avoid squeezes traders may not have more than 150 contracts long or short on the Friday in the first week of the expiration month.

Table I.C.7.4: Commodity contracts and daily limits on price movements

CONTRACT	EXCHANGE	LIMIT	PRICE	TICK MOVE	CONTRACT SIZE	UNITS
Cocoa	CSCE	88	1995:1995	\$10.00	10 tonnes	\$/tonne
Hogs (lean index)	CME	150	40.15:4015	\$4.00	40,000 lbs	c/lb
Pork bellies, frozen	CME	200	52.27:5227	\$4.00	40,000 lbs	c/lb
Cotton #2	NYCE	300	47.69:4769	\$5.00	50,000 lbs	c/lb
Copper, hi-grade RTH ²⁶	COMEX	2000	105.50:10550	\$2.50	25,000 lbs	c/lb
Corn RTH	CBT	120	162.25:1622	\$6.25	5,000 BU	c/BU
Coffee	CSCE	600	115.45:11545	\$3.75	37,500 lbs	c/lb
Oats RTH	CBT	100	143.75:1436	\$6.25	5,000 BU	c/BU
Orange juice	NYCE	500	125.25:12525	\$1.50	15,000 lbs	c/lb
Platinum	NYMEX	250	508.10:5081	\$5.00	50 troy oz	\$/oz
Silver RTH	COMEX	1500	571.5:5715	\$5.00	5,000 troy oz	c/oz
SoybeanS RTH	CBT	300	451.5:4514	\$6.25	5,000 BU	c/BU
Soybean meal RTH	CBT	100	134.80:1348	\$10.00	100 tons	\$/ton
Soybean oil RTH	CBT	100	16.65:1665	\$6.00	60,000 lbs	c/lb
World sugar #11	CSCE	50	8.98:898	\$11.20	112,000 lbs	c/lb
Wheat RTH	CBT	200	282.5:2824	\$6.25	5,000 BU	c/BU
Kansas City wheat	KCBT	250	274.5:2744	\$6.25	5,000 BU	c/BU
Lumber	CME	100	174.90:1749	\$8.00	80,000 BF	\$/1000 BF
Gold	CBT	500	458.50:4585	\$10.00	100 ounces	\$/oz
Stocker cattle	CME		78.85:7885	\$2.50	25,000 lbs	c/lb
Gold RTH	COMEX	750	387.50:3875	\$10.00	100 troy oz	\$/oz
Feeder cattle	CME	150	63.70:6370	\$5.00	50,000 lbs	c/lb
Silver	CBT	1000	767.0:7670	\$5.00	5000 troy oz	c/oz
Domestic sugar #14	CSCE	50	19.00:1900	\$11.20	112,000 lbs	c/lb
Oats	WCE	500	84.50:8450	CAD 0.20	20 tonnes	CAD/tonne
Rapeseed (canola)	WCE	100	214.10:2141	CAD 2.00	20 tonnes	CAD/tonne
Flaxseed	WCE	100	192.00:1920	CAD 2.00	20 tonnes	CAD/tonne
Feed wheat (domestic)	WCE	500	90.40:9040	CAD 0.20	20 tonnes	CAD/tonne

²⁶ RTH stands for Regular Trading Hours, meaning 'trading hours on the exchange'. Once the exchange closes, many contracts are traded CT (Trading Cycle) through electronic networks such as Globex or Access. See for example, <http://www.cme.com/trd/calhrs/tradehours3497.html>, which defines RTH and CT hours.

I.C.7.5 Risk Management at the Commodity Trading Desk

This section explains the first golden rule for commodities: that one must decompose risks into those associated with an outright spot position and those defined by the borrowing cost of commodities. Such decomposition can greatly ease the computational burden when calculating value-at-risk (VaR) for a portfolio of commodities. Furthermore, it enables the risk manager to assess where the concentration of risk is based. Examples of VaR calculations for commodity portfolios are provided here (see Chapter III.A.2 for a general discussion of VaR). We show that for a typical market making commodity portfolio in cash and derivatives, most risk comes from volatility in the convenience yield. The portfolio includes all the hedges of the trading desk, and of course if VaR limits or other limits are imposed then these will also affect the VaR of the observed portfolio. Generally, traders hedge out their spot exposure – the reason is that the market is relatively liquid but long-dated interest-rate risk is more expensive to hedge, and thus traders must balance transactions costs with the uncertainty of the unhedged exposures (see, for example, Lawrence and Robinson, 1995a, 1995b, 1996).

The no-arbitrage condition (equation (I.C.7.2)) in an efficient market, where the expected spot price is the future price, and when the storage costs are constant, gives

$$\sigma_F = \sigma_S + \Delta r - \Delta q, \tag{I.C.7.4}$$

where σ_F is the percentage change in the futures price, σ_S is the percentage change in the spot price and Δr and Δq denote the change in the interest rate and the convenience yield, respectively. The term $\Delta r - \Delta q$ is the borrowing cost (net of the immediacy premium) and we here refer to the percentage price changes as the volatility. An example of the decomposition of variance as measured by a standard deviation, for aluminium and gold futures at 3 months and 27 months, is given in Table I.C.7.5.

Table I.C.7.5: Decomposition of risk in computing VAR

Commodity	Decomposition of Volatility			Decomposition of Net Borrowing Volatility		
	Forward Volatility	Spot Volatility	Net Borrowing Cost	Net Borrowing Costs Volatility	Convenience Yield	Interest Rate
Alum 3M	17.54%	19.10%	-1.56%	2.48%	2.44%	0.05%
Alum 27M	16.67%	8.48%	8.19%	19.07%	19.40%	-0.33%
Gold 3M	4.86%	5.06%	-0.20%	1.05%	0.35%	0.69%
Gold 27M	4.82%	4.81%	0.00%	1.87%	0.62%	1.25%

Several features stand out. A position in a forward is analogous to a portfolio of two risk factors: spot plus borrowing costs. For shorter-dated base commodities, the spot affect dominates. In aluminium three-month forwards the outright volatility is 17.54%, indeed the borrowing costs reduce overall the risk. In contrast, aluminium long-dated 27-month contract risk is shared equally between net spot volatility of 8.48% and net borrowing cost of 8.19%. In gold the spot rate dominates the overall contribution of, risk with borrowing costs contributing little.

The immediacy premium is related to the volatility of net borrowing costs. The right-hand side of the table shows that the interest rate affect dominates the immediacy premium for gold. However, it is the convenience yield that determines the (much higher) immediacy premium for aluminium. Lawrence (2003) has analysed the key reason why base metals tend to have more volatile convenience yields than gold. Gold essentially has extremely large above-ground stocks that can readily be converted into inputs to meet industrial demand. The law of diminishing marginal utility from liquid commodities such as gold leads to a small and less volatile convenience yield.

We have selected a real portfolio of aluminium, copper, lead, nickel, tin and zinc to illustrate these ideas. Table I.C.7.6 describes the spot and forward positions (in £000) of a real portfolio of a major trading bank.

Table I.C.7.6: A portfolio of commodities

Tenor	Alum	Alum Alloy	Copper A	Lead	Nickel	Tin	Zinc
Spot	98.9	9.1	4.8	9.8	8.6	2.6	151.3
1 wk	-15.1	-5	-10.3	-5.3	8.1	-1.8	-42.8
2 wk	-4	-1.6	-0.5	-0.3	-8.9	0.2	-65.1
1 m	-23.1	-2.1	-8.5	2.4	-4.8	0.1	-7.6
3 m	15.7	0.2	-4.2	-6	-2	-0.7	-6.9
6 m	-21	0.5	-4.7	-0.8	-0.7	-0.1	-19.3
12 m	-17.4	-0.9	14.9	-0.3	0	0	-8.5
15 m	1.5	0	6.5	0	0	0	-0.7
18 m	-3.2	0	0.6	0	0	0	-0.1
2 yr	-1.5	0	-4.8	0	0	0	0
27m	0.7	0	1.2	0	0	0	0
36 m	0	0	0.6	0	0	0	0

The portfolio is a diversified portfolio of long spot positions with hedges (short forwards) mainly concentrated in the short-dated maturities but extending out to 3 years. In order to compute the VaR of the portfolio one could use a covariance model, but this would be a cumbersome process: we would need over 2500 parameters (correlations and variances) to estimate the VaR. The

alternative approach is to decompose the forward risks into spot risk and the risk associated with carrying costs, as demonstrated above. To compute the overall spot position in each commodity, each forward must be discounted to an equivalent spot position (using equation (I.C.7.3)). The overall spot positions are computed in Table I.C.7.7.

In this table, the annual volatility in the centre column is translated into a daily VaR based on a normal distribution at the 98% confidence level, using the method described in Chapter III.A.2. Of the individual VaRs for each commodity (termed ‘gross’ VaR in the second last column) the largest is aluminium with a VaR of £181,450, and the second largest is copper with a VaR of £95,000.

Table I.C.7.7: Overall spot exposures, volatility and VaR

	Physical Position (000s tonnes)	Value Physical (£millions)	Volatility (annual %)	Individual (‘Gross’) VaR (£000)	Contribution to Total VaR (£000)
Alum	0.16	0.17	29%	181.45	163.59
Alum Alloy	0.28	0.26	15%	7.53	5.29
Copper Grade A	−4.35	−8.47	14.90%	95.76	−21.21
Lead	−0.57	−0.27	24.37%	7.19	−2.12
Nickel	0.1	0.54	23.41%	43.6	25.37
Tin	0.32	1.31	24.47%	43.34	26.33
Zinc	0.31	0.21	20.55%	64.23	37.3

The most interesting column, though, is the net contribution of each position to the overall risk. Diversification effects are captured using the correlation between spot returns. Aluminium risk declines to £163,000 and the short copper position with an individual VaR of £95,000 reduces overall exposure by £21,000 since copper and aluminium have a positive correlation of 0.54.

By adding up the net risk contribution in the final column of Table I.C.7.7 we compute the total VaR of the spot positions as approximately £230,000. The aggregate risk of £440,000 is the sum of the individual VaRs in the second last column of Table I.C.7.7. This represents the risk if each position moved adversely against us. The risk reduction due to diversification is £440,000 – £230,000 = £210,000. This demonstrates that setting limits on each commodity can be too constraining and will exaggerate VaR. Indeed, diversification reduces risk by almost 50%.

This representation of spot risk assumes that the commodity yields do not change. To get a handle on the borrowing risk, Table I.C.7.8 analyses the borrowing cost sensitivities of each commodity at each tenor. The report first computes the sensitivity of each commodity at each

tenor to a 100 basis point rise in commodity yields. These sensitivities are shown in the top part of the table. The sum of these sensitivities over all tenors gives the ‘bull–bear’ sensitivity: this is positive for a bear sensitivity and negative for a bull sensitivity. For example, overall we have a ‘bull’ position of £436,000 in aluminium.

Table I.C.7.8: Base metals borrowing/lending premium report

Change in Value for							
1% in							
Borrowing/Lending	Aluminium		Copper				
Premium (£'0000s)	Aluminium	Alloy	Grade A	Lead	Nickel	Tin	Zinc
Spot	-3	0	0	0	1	0	3
1 wk	-3	-1	-4	0	8	-1	-5
2 wk	-2	-1	0	0	-18	0	-17
1 month	-21	-2	-14	1	-22	0	-4
3 month	-44	1	-19	-7	-27	-7	-12
6 month	-119	3	-39	-2	-20	-1	-67
12 month	-202	-9	239	-1	-2	1	-60
15 month	22	0	128	0	0	0	-6
18 month	-55	0	15	0	0	0	-1
2 year	-34	0	-146	0	0	0	0
27 month	19	0	39	0	0	0	0
36 month	0	0	28	0	0	0	0
Bull–Bear Sensitivity	-436	-9	227	-10	-79	-8	-170
Adverse Risk	454	16	1215	37	88	18	184
Net Parallel Risk	301	4	443	25	102	8	143
Net Non-parallel Risk	59	5	-125	8	-66	5	9
Daily Value-at-Risk	360	9	319	32	36	12	152
Portfolio (£millions)							
Adverse Risk	2.01						
<i>Less Diversification Effect</i>	1.35						
98% Daily Value-at-Risk							
(Normality)	0.46						
<i>Add Non-Normality Effect</i>	0.18						
98% Daily Value-at-Risk	0.64						

A bull position in commodities is identical to selling spot and hedging with a forward position. With a bull position a rise of 100 basis points across the whole yield curve would lead to a loss. This rise in the commodity yield could come about from either a rise in the interest rate, a fall in the convenience yield or a hike in storage costs.

Table I.C.7.8 also provides data to traders on shifts in the term structure of interest rates. In the example shown most tenors have negative sensitivities. Now, using a simulation technique, we compute the borrowing risk VaR for each commodity. For example, at the bottom of the aluminium column, the VaR is £360,000 and this is composed of £300,000 in parallel shifts in the aluminium yield curve and £59,000 in non-parallel shifts. The above numbers are estimates of the net contribution of risk to the borrowing cost portfolio, assuming that the spot price is held constant.

Table I.C.7.9: Aggregate borrowing cost risk

Gross Risk	£2,000,000
Diversification	-£1,440,000
VaR	£460,000

Table I.C.7.9 shows the aggregate borrowing cost risk of the example portfolio. By adding up the gross VaR of each commodity we estimate the adverse situation where all yield curves shift adversely against us. This is estimated at £2 million. However, since the correlations between yield curves across commodities are generally low, the estimate of diversification at the 98% confidence interval is £1.44m. Thus the VaR of borrowing costs is estimated at £460,000.

In Table I.C.7.10 we produce the overall VaR and decomposition of the portfolio by commodity into spot and carrying cost exposure as well as an estimate of the aggregate VaR. For the portfolio as a whole we noted that the borrowing rate VaR is about £460,000 and spot commodity VaR is £230,000. Most of the commodity risk is entirely due to the ‘convenience yield’ risk.

For most commodities we note how the desk has much larger borrowing cost risk than spot risk. For example, in copper the spot VaR is £97,000 and carrying cost VaR is £280,000. This need not be the case, but it is something that typically occurs at trading desks due to hedging of spot exposure but inability to hedge term structure risk (particularly long-dated forwards) due to illiquidity and high transactions costs.

The demand for immediacy is a core risk factor contributing significantly to overall risk. In a simulation of the above portfolio, we found that if the convenience yield were held constant, the VaR of the portfolio would fall by 75%. A bank that is running a short metals position hedged by a longer-dated forward contract can be subject to a severe market squeeze if the market suddenly goes into backwardation. Stress testing and empirical scrutiny of borrowing rate (the forward–

spot contango) behaviour take on considerable importance for the risk assessment of a commodity portfolio.

Table I.C.7.10: Decomposition of VaR

(£000s)	Aluminium	Aluminium Allo	Copper Grade A	Lead	Nickel	Tin	Zinc
Spot Risk	181.5	7.5	96.8	7.2	43.6	43.3	64.2
Borrowing Rate Risk	367.4	9	280.3	31.1	33.2	12.1	143
Daily VaR (£000s)	409.8	11.8	296.6	31.9	54.8	45	157
Portfolio Risk (£Millions)		Decomposition of Risk (£Millions)					
Adverse Risk	1.98	Spot Commodity Risk	0.23				
<i>Less Diversification</i>	1.46	Borrowing Rate Risk	0.46				
98% Daily VaR (Normality)	0.51	Daily Value-at- Risk	0.51				
<i>Add Non-normality Effect</i>	0.26						
98% Daily VaR	0.77						
Limit	2						

If we added up all the individual VaRs this would total £1.98 million. But it appears that all these risks are empirically quite independent, so that diversification would reduce the exposure to £510,000. But this only applies if the portfolio returns are normally distributed. Re-estimating the entire portfolio under a nonparametric distribution, we do find significant non-normality. The VaR now increases by 25% to £710,000. This suggests that non-normality is absolutely vital in estimating VaR for commodity portfolios.²⁷

I.C.7.6 The Distribution of Commodity Returns

We have shown that commodity prices have been subject to sudden unexpected shocks such as harvest failures or production delays, leading to an increase in the convenience yield. We have further explored some illustrations of squeezes or potential squeezes on physical deliveries into forward or futures contracts. Consequently, we were not at all surprised to observe that normal distributions were inconsistent with empirical observation in the VaRs estimated in Section I.C.7.5. We further explore here the characteristics of the observed distributions.

²⁷ See Dusak (1973), who provides a methodology for testing whether or not commodity future prices are non-normally distributed. For soy beans and other softs she rejects the normality assumption. For the risk manager this suggests that to estimate economic capital for a commodity desk using a normal distribution assumption could seriously understate risk.

I.C.7.6.1 Evidence of Non-normality

Table I.C.7.11 presents summary statistics of quarterly data over the period 1982–2001 for financial instruments and some commodities. The series are the three-month Treasury bill rate, the S&P index, bonds, gold, the CRB index, aluminium copper, lead, zinc, oil and silver. All data have been deflated by the wholesale price index so that we can examine real rates of return.

The commodities all have *negative* rates of return over this period. But, amazingly, the average annual volatility of gold over the period is about the same as that of the S&P index and Dow Jones.

Table I.C.7.11: Sample moments of commodity return distributions

Asset	Mean	Volatility	Skewness	Kurtosis
R3M	4.38	5.16	0.39	0.21
RSP	6.78	32.21	−0.54	0.95
RBOND	5.5	21.32	0.12	1.18
RGOLD	−1.36	34.49	0.7	2.14
RCRB	−3.03	19.95	−0.14	0.09
RALUM	−1.63	45.59	0.11	2.1
RCOPPER	−2.71	46.07	0.67	2.22
RLEAD	−3.3	56.89	−0.09	1.23
RZINC	−3.21	44.71	0.03	0.19
RWTI	−0.98	59.76	−0.24	7.91
RSILVER	−2.8	53.91	0.56	2.95

Commodity returns are also highly non-normally distributed. The last two columns of Table I.C.7.11 show the skewness and excess kurtosis estimates, again based on quarterly data. At this frequency the financial asset returns are near to normally distributed – the 1987 outlier being the main cause of non-normality in the S&P index. However, aluminium, gold, silver, copper and oil all have highly leptokurtic distributions. Hence their risk management requires the use of non-normal VaR models (see Chapter III.A.3).

I.C.7.6.2 What Drives Commodity Prices?

Lawrence (2003) has explored the key drivers of commodity prices. Using regressions of commodity returns on gross domestic product, inflation, monetary growth, short- and long-term yields as proxies of the business cycle, he examines how prices fluctuate with the business cycle for each commodity. Consequently, he categorises commodities into two classes: those that are correlated with the cycle (through one or more variables) and those that are not. Most commodity prices are related to a business cycle, through varying channels. Only zinc and gold appear to have no systematic risk. Any commodity that does not have excessive above-ground stocks relative to production flows, such as aluminium or oil, will tend to be affected by cyclical

movements. The astute risk manager should perform stress tests and scenario analysis (see Chapter III.A.4) and be aware that estimates of correlations vary dramatically over the cycle (see Chapter II.A.3).

I.C.7.7 Conclusions

This chapter covers the markets for and the transaction characteristics of different types of commodities. As a risk manager it is important to recognise these characteristics, which focus on the delivery and settlement mechanisms for heterogeneous commodities. The key arbitrage equation which links spot prices with forward prices is the commodity equivalent of covered interest arbitrage in foreign exchange or interest parity in bonds. Unlike other markets, the arbitrage equation for commodities contains a convenience yield which reflects the importance that is sometimes placed on immediate access to supply. This feature of commodity markets is no doubt related to the importance of commodities as factors of production and possible delays in supply/shipping. It is the presence of this convenience yield, and its variability, that make commodity risk management unique. Failure to properly appreciate these aspects of commodity markets can have disastrous consequences.

We have introduced the reader to the risk management of a commodity trading desk. We have shown how any forward commodity position can be decomposed into an outright spot position and carrying cost position. We have presented the reader with examples of the key value-at-risk and sensitivity reports that are produced by commodity trading desks. These reports highlight the importance of diversification for portfolios of commodities, a characteristic held in common with trading portfolios in other markets (such as stocks, bonds and currencies). Unlike these other markets, the main component of risk for a commodities portfolio is typically not spot risk, but changes in the cost of carry. The markets are also subject to extreme movements due to squeezes or corners. We showed two examples of the golden age manipulators in the nineteenth century. As a consequence, the CFTC and the exchanges responded by setting limits on maximum price fluctuations, maximum position size and criminal legal action. These still did not terrify the Hunts, who allegedly created the greatest corner in 1979 and even had the central bankers lending them money.

Commodity risk managers should also consider the unique characteristics of the distribution of commodity returns. Compared with financial assets, commodities typically have high volatility and low returns. Many also exhibit non-normality, with positive excess kurtosis. Examining the key risk factors that drive the prices of commodities, we conclude that the business cycle only affects commodities that have small above-ground stocks relative to production. In such cases an

immediacy premium can arise due to sudden unexpected shortages in input supply. This can drive markets into backwardation. In contrast, when there is a ready supply of above-ground stocks that can be converted into production inputs, as in the case of gold, the business cycle has little effect on pricing and there is less chance of dramatic hikes in spot prices over and above forward prices.

References

Dusak, K (1973) Futures trading and investor returns: An investigation of commodity market risk premiums. *Journal of Political Economy*, 81, pp. 1387–1406. Reprinted in Telser (2000, pp. 597–617).

Geisst, C R (2002) *Wheels of Fortune. The History of Speculation from Scandal to Respectability*. Hoboken, NJ: Wiley.

Hicks, J R (1946) *Value and Capital*, 2nd edition. Oxford: Oxford University Press.

Kaldor, N (1939) Speculation and Economic Stability. *Review of Economic Studies*, 7(1), pp. 1–27. Reprinted in Telser (2000, pp. 53–87).

Keynes, J M (1930) *A Treatise on Money*, Volume 2. London: Macmillan.

Lawrence, C (2003) Why is gold different from all other assets? An empirical investigation. World Gold Council.

Lawrence, C, and Robinson, G (1995a) Handbook of market risk management. Unpublished manuscript, Barclays Bank.

Lawrence, C, and Robinson, G (1995b) Value at risk: addressing liquidity and volatility risks. *Capital Market Strategies*, no. 7 (November).

Lawrence, C, and Robinson, G (1996) Incorporating Liquidity into the Risk-Measurement Framework *Financial Derivatives and Risk Management*, no. 6 (June).

Reuters Limited (2000) *An Introduction to the Commodities, Energy and Transport Markets*. Chichester: Wiley.

Sharpe, W (1964) Capital asset prices: A treaty of market equilibrium under conditions of risk. *Journal of Finance*, 19, pp. 425–442.

Sikorzewski, W (2001) Corners in the commodity futures markets in the XIXth century. Unpublished manuscript, University of Caen, France.

Telser, L G (1958) Futures trading and the storage of wheat and cotton. *Journal of Political Economy*, 66, pp. 233–255. Reprinted in Telser (2000) pp. 119–148.

Telser, L G (2000) *Classic Futures-Lessons from the Past for the Electronic Age*. London: Risk Publications.

Telser L.G and Higinbotham, H N (1977) Organised futures markets: costs and benefits. *Journal of Political Economy*, 85(5), pp. 969–1000. Reprinted in Telser (2000, pp. 321–353).

Williams J (1986) *Economics of Futures Markets*. Cambridge: Cambridge University Press.

I.C.8 The Energy Markets

Peter C. Fusaro¹

I.C.8.1 Introduction

Energy trading began in 1978 with the first oil futures contract on the New York Mercantile Exchange (NYMEX). During the 1980s and 1990s, NYMEX and the International Petroleum Exchange (IPE) successfully launched futures contracts for oil and gas futures trading. These successful energy futures exchanges have survived the trading debacles of recent years, of which Enron was the most notable. Oil companies and financial houses now provide the necessary trading liquidity through market-making on both the established government-regulated futures exchanges and over-the-counter (OTC) energy derivatives markets that can clear on the futures exchanges. They have considerable skill in the management of financial energy risks and the risks in the emerging global environmental markets.

This chapter introduces the energy markets from a risk management perspective. Section I.C.8.2 provides an initial overview of the market, its products and its risks. Section I.C.8.3 explains how risk management is conducted on energy exchanges, while Section I.C.8.4 covers the OTC markets. In both cases a global perspective is given. Emerging energy markets are the subject of Section I.C.8.5, including brief coverage of coal trading, weather derivatives, green trading and freight swaps. Section I.C.8.6 looks to the future of energy trading, while Section I.C.8.7 concludes.

I.C.8.2 Market Overview

Crude oil and petroleum products have particularly active markets. Daily physical oil consumption is over 80 million barrels per day and approaches \$1 trillion in annual trade, which entices many active hedgers and speculators. Crude oil and petroleum products are now traded 24 hours per day every business day in both the physical and paper (financial) markets. Options and OTC oil price swaps are also well developed. The majority of financial oil trading still takes place in the United States and Europe. Although Asian energy trading markets in general are less developed, that is the region with the highest oil consumption growth rates. Natural gas has had a viable energy futures contract in North America since April 1990, and in Europe since 1999.

Electric power futures and OTC contracts have proliferated around the world since the 1990s but, frankly, these markets are much smaller than other markets, mainly due to the high price risk

¹ Chairman, Global Change Associates Inc., New York, NY (www.global-change.com).

and the inability to store electricity. Moreover, there have been many failed electricity futures contracts in the USA and electricity futures exchanges in Europe. While paper market trading for oil and gas has grown considerably during the past decade, on both established futures exchanges and the OTC forward markets, electricity paper trading is still in its infancy. Electricity deregulation has driven the commoditisation process whereby electricity becomes a fungible commodity, as it is in the Nord Pool market in Scandinavia (perhaps the best example of a working electricity commodity market in the world). There is convergence of both gas and electric prices that has accelerated much more on the physical side of the market than the financial trading of power. In fact, the close relationship between natural gas and electric power markets cannot be understated. However, power is a more demanding market, as it is a next hour, next day, next week, and next month business. Power marketers and traders provide greater efficiency by buying and selling power and transmissions capacity. Electric power is a market that never closes, where prices change hourly, half-hourly or quarter-hourly. It is the most volatile commodity ever created and therefore its financial markets are significantly smaller than the oil and gas markets. While there is much attraction in trading electricity, the fact is that it is a very hard risk to manage, with fuel inputs on one side of the equation and electricity outputs on the other side. Adding to this risk complexity is the fact that electric utilities are the main emitters of greenhouse gases and other air emissions. This will add costs and more trading complexity to trading electric power in the future as the environmental financial markets become better established throughout the world.

Today we are still only hedging about half of the global commodity price exposure of the physical energy markets. To put this statement into some perspective, the annualised notional value of all energy derivatives is about \$2 trillion, compared to a physical energy market of \$4 trillion annually. Commodities usually trade at least 6 and up to 20 times the physical market. Foreign exchange and government bonds today are over \$190 trillion in notional value. Consequently, financial energy trading is still in its infancy. Ironically, most companies still do not hedge their energy price risk exposures. However, the new market drivers for energy hedging and active risk management are growing because of market liberalisation, competitive markets, globalisation, privatisation of energy, and the Internet, which is facilitating electronic energy trading. Increased price volatility due to the more active trading of hedge funds and investment banks makes this sector ripe for financial change.

I.C.8.2.1 The Products

The energy complex trades the following products on established futures exchanges, OTC markets and the Internet:

- crude oil

- gasoline
- naphtha
- gasoil
- jet fuel
- home heating oil
- residual fuel oil
- bunker fuels
- freight-rate swaps
- natural gas
- electricity
- liquefied natural gas (LNG)
- petrochemicals
- coal
- emissions such as sulphur dioxide and nitrous oxides
- greenhouse gases
- renewable energy credits
- negawatts (value of energy efficiency)

Some of these products are quite illiquid (like greenhouse gases or negawatts) while others (like crude oil in the USA and Europe as well as North American natural gas) are very well established. The mature energy commodity markets have more sophisticated financial instruments and recently have attracted the active participation of many hedge funds which are extending their trading platform from energy equities into energy commodities as they seek higher returns.

I.C.8.2.2 The Risks

Energy commodities are subject to numerous risks, including credit or counterparty risk, liquidity risk, event risk, cash-flow risk, basis risk, legal and regulatory risk, operational risks, tax risk and most evidently geopolitical and weather risks. There are also tremendous variations over time in many energy markets. The weather (seasonal) impacts supply and demand so that risks increase in the mid-summer and winter seasons as more energy is required for heating, cooling, and transportation.

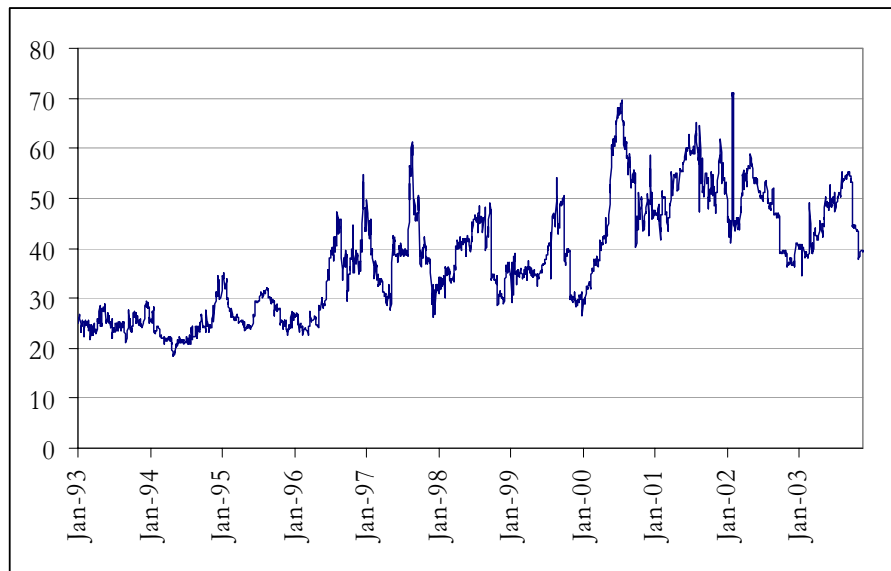
Of all the different types of risks that affect the energy markets, market risk is still pre-eminent. Price volatility is caused by fundamental factors such as supply/demand, and weather and financial factors such as technical trading, speculators, and market imperfections. These factors

are very well defined in energy markets and as a result they are the most volatile commodity markets ever created.

Figure I.C.8.1: Implied volatility, crude oil

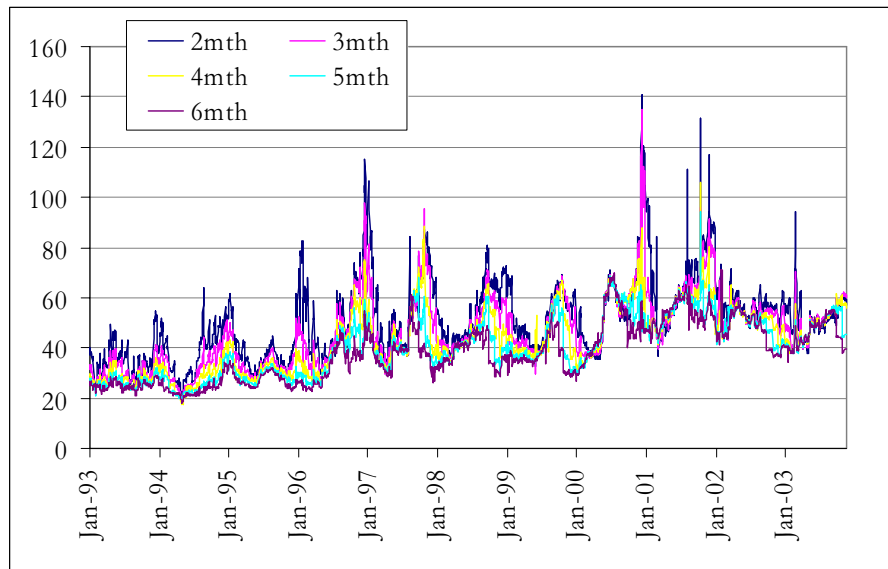


Figure I.C.8.2: Implied volatility, natural gas



Figures I.C.8.1 and I.C.8.2 show the implied volatility of six-month options on light sweet crude oil and natural gas, both from NYMEX for the period from 4 January 1993 until 20 November 2003. Whilst shorter-term options on crude oil have volatilities in the same range as the six-month implied volatility shown in Figure I.C.8.1, the prices of very short-term natural gas options are much affected by supply shortages, often during cold winter months. For instance, the two-month implied volatility has exceeded 100% several times during the last decade, as shown in Figure I.C.8.3. Electric power volatility sometimes reaches 1000% due to supply shortages and weather-related influences. Even coal trading has been demonstrating unusually high volatility during the past year, of 50% to 100%. Clearly, energy markets are the most volatile financial markets, and consequently risk management has become a core competency and fiduciary responsibility for many energy companies.

Figure I.C.8.3: Implied volatility term structure, natural gas



I.C.8.2.3 Developing a Cash Market

The most interesting and complex factor that differentiates energy commodity trading from forex and other financial instruments is its need for a viable cash market to enable trading. Today, we have viable cash markets for the following energy commodities: crude oil, heating oil (also called gasoil in Europe), gasoline, propane, residual fuel oil, ship bunkers, natural gas, coal and electricity. Energy market development always begins with opaque prices, little trading, poor liquidity, few participants, wide arbitrage opportunities, fat margins and tremendous inefficiency. At this point in market development, deal flow is dominated by OTC brokers. As the cash

market develops spot market price discovery there is more deal flow.² Eventually both OTC forward contracts and exchange-traded futures start to trade. There is no special time period for all markets to develop, but the evolutionary process is the same. Although energy price management is still in its early stages of development compared to the more developed and financially sophisticated markets of interest rates and foreign currencies, the future is indeed very bright.

I.C.8.3 Energy Futures Markets

A standardised energy futures contract always has the following characteristics. It has an underlying physical commodity or price index upon which the energy futures contract is based. There is a certain size for the amount of the underlying item covered by each futures contract. There is a predetermined and specified time given in months for which contracts can be traded. There is an expiration date. Finally, there is a specified grade or quality and delivery location for oil and coal futures contracts. Whereas oil varies by grade/quality, natural gas (methane) and electricity are more homogeneous commodities, obviating the need for the grade/quality to be specified in the contract. The settlement mechanism can be either physical delivery of the underlying item or cash payment. The trend in energy futures has been towards cash settlement, but in reality the only liquid futures contract that goes to cash settlement is the IPE Brent crude oil contract.

Partly because futures markets provide the opportunity for leveraged investments, they attract large pools of risk capital. As a result, futures markets are among the most liquid of all global financial markets, providing low transaction costs and ease of entry and exit. This fosters their use by a wide range of businesses and investors wishing to manage price risks. Due to the continued geopolitical problems over oil supply, futures exchanges are currently experiencing unprecedented growth in their oil futures contracts. Futures markets also are government-regulated which guarantees performance.

I.C.8.3.1 The Exchanges

Futures markets have been used by traders in commodities for hundreds of years, beginning with the trading of rice futures in Osaka, Japan, in the late 1600s. NYMEX, the world's largest regulated energy futures exchange, started life in 1872 as the Butter and Cheese Exchange of New York before being renamed 10 years later. Today, NYMEX is the largest energy futures exchange in the world, having introduced its first energy futures contract for home heating oil in 1978, and trades oil, gas, electric power and coal contracts. Other oil and gas futures markets

² The spot market is a commodity market for sale and delivery of energy. Spot markets exist for oil, natural gas, electricity, and coal. 'Price discovery' refers to the process of determination of market prices through the interactions of buyers and sellers in the marketplace.

include the IPE in London, which trades oil and gas contracts, and the Tokyo Commodity Exchange in Tokyo, which trades small oil futures contracts. The Singapore Exchange used to trade fuel oil futures contracts in the past and now has relationships with other energy futures exchanges for electronic trading. The Shanghai Futures Exchange (SHFE) launched a fuel oil futures contract in August 2004 and also has a number of linkages to NYMEX, the IPE and the Singapore Exchange.

There are a number of electricity exchanges which trade both physical spot electricity and electricity futures, but the three largest are the UK Power Exchange (UKPX), NordPool in Scandinavia and the European Energy Exchange (EEX) in Germany. NordPool (www.nordpool.com) is the most liquid electricity trading market in the world. It trades OTC, bilateral, cleared OTC, and physical forward electricity contracts. UKPX (www.ukpx.com) provides a market for trading in both spot and futures contracts in electricity. It also acts as a clearing house for OTC transactions. EEX (www.eex.de) trades both physical spot electricity and electricity futures and is the consolidation of two German exchanges. Then there are several electronic energy platforms, the IntercontinentalExchange™ (ICE) of the USA being the largest. ICE bought IPE in June 2002 and is now the second largest energy exchange in the world; it has launched the Interchange, which trades the IPE during floor trading hours. The ICE trades many oil, gas and power OTC contracts in North America and Europe, but its primary growth in recent years has been in Asia. There is another electronic exchange called the Natural Gas Exchange in Calgary, Canada, which has been profitable for many years and trades a western Canadian natural gas contract.

To round out the remaining energy futures exchanges, there is the Amsterdam Exchange (www.apx.nl) in the Netherlands which engages in day-ahead power trading, Powernext Paris (www.powernext.fr) in France which also engages in day-ahead power trading, the Austrian Power Exchange (www.exaa.at) which has limited activity, the Warsaw Power Exchange (www.polpx.pl) which trades day-ahead physical electric power, and the Spanish Power Market (www.cne.es) in Madrid which trades physical power in Spain. None of these exchanges trade high volumes. The problem in the European Union is that all countries have developed electronic exchanges independently and have not figured out that the Internet is borderless, and that all these exchanges are doomed to failure unless they link to or are subsumed by other exchanges.

I.C.8.3.2 The Contracts

The major oil futures contracts are NYMEX light sweet crude oil futures, NYMEX heating oil futures, NYMEX gasoline futures, IPE Brent crude oil futures, and IPE gasoil futures. Natural

gas futures contracts are only traded on the NYMEX and IPE. The NYMEX Henry Hub natural gas futures contract is considered the global benchmark for gas trading, and recently has been used for LNG hedging purposes. It was launched in 1990. The IPE's natural gas futures contract is more regionalised for the UK's National Balancing Point (price reference) and was launched in 1997.

Example I.C.8.1: Contract specifications – Light, sweet crude oil futures

Trading Unit: 1000 US barrels (42,000 gallons).

Price Quotation: US dollars and cents per barrel.

Trading Hours (all times are New York time): Open outcry trading is conducted from 10:00 a.m. until 2:30 p.m. After hours futures trading is conducted via the NYMEX ACCESS® Internet-based trading platform beginning at 3:15 p.m. on Mondays to Thursdays and concluding at 9:30 a.m. the following day. On Sundays, the session begins at 7:00 p.m.

Trading Months: Thirty consecutive months plus long-dated futures initially listed 36, 48, 60, 72 and 84 months prior to delivery. Additionally, trading can be executed at an average differential to the previous day's settlement prices for periods of 2–30 consecutive months in a single transaction. These calendar strips are executed during open outcry trading hours.

Minimum Price Fluctuation: \$0.01 (1¢) per barrel (\$10.00 per contract).

Maximum Daily Price Fluctuation: \$10.00 per barrel (\$10,000 per contract) for all months. If any contract is traded, bid, or offered at the limit for five minutes, trading is halted for five minutes. When trading resumes, the limit is expanded by \$10.00 per barrel in either direction. If another halt were triggered, the market would continue to be expanded by \$10.00 per barrel in either direction after each successive five-minute trading halt. There will be no maximum price fluctuation limits during any one trading session.

Last Trading Day: Trading terminates at the close of business on the third business day prior to the 25th calendar day of the month preceding the delivery month. If the 25th calendar day of the month is a non-business day, trading shall cease on the third business day prior to the business day preceding the 25th calendar day.

Settlement Type: Physical.

Delivery: Free on board seller's facility, Cushing, Oklahoma, at any pipeline or storage facility with pipeline access to TEPPCO, Cushing storage, or Equilon Pipeline Co., by in-tank transfer, in-line transfer, book-out, or inter-facility transfer (pumpover).

Delivery Period: All deliveries are rateable over the course of the month and must be initiated on or after the first calendar day and completed by the last calendar day of the delivery month.

Alternative Delivery Procedure (ADP): An ADP is available to buyers and sellers who have been matched by the Exchange subsequent to the termination of trading in the spot month contract. If buyer and seller agree to consummate delivery under terms different from those prescribed in the

contract specifications, they may proceed on that basis after submitting a notice of their intention to the Exchange.

Exchange of Futures for Physicals (EFP): The commercial buyer or seller may exchange a futures position for a physical position of equal quantity by submitting a notice to the Exchange. EFPs may be used to either initiate or liquidate a futures position.

Deliverable Grades: Specific domestic crudes with 0.42% sulphur by weight or less, not less than 37° API gravity or more than 42° API gravity. The following domestic crude streams are deliverable: West Texas Intermediate, Low Sweet Mix, New Mexican Sweet, North Texas Sweet, Oklahoma Sweet, and South Texas Sweet. Specific foreign crudes of not less than 34° API or more than 42° API. The following foreign streams are deliverable: UK Brent and Forties, and Norwegian Oseberg Blend, for which the seller shall receive a 55¢ per barrel discount below the final settlement price; Nigerian Bonny Light and Colombian Cusiana are delivered at 15¢ premiums; and Nigerian Qua Iboe is delivered at a 5¢ premium.

Inspection: Inspection shall be conducted in accordance with pipeline practices. A buyer or seller may appoint an inspector to inspect the quality of oil delivered. However, the buyer or seller who requests the inspection will bear its costs and will notify the other party to the transaction that the inspection will occur.

Position Accountability Levels and Limits: Any one month/all months: 20,000 net futures, but not to exceed 1000 in the last three days of trading in the spot month.

Margin Requirements: Margins are required for open futures positions.

Trading Symbol: CL

Source: New York Mercantile Exchange

Electricity futures are traded on the NYMEX for the eastern United States in its PJM³ and NYISO⁴ electricity futures contracts (among others), UKPX for the UK power markets, NordPool for the Scandinavian power markets and EEX for the German power markets. There are many other minor electricity exchanges in Europe, including some in France, Austria, Poland and the Netherlands. The IPE is preparing to launch its second attempt at a successful electricity futures contract.

It should be stated that most energy futures contracts fail. It takes several success factors to make an energy futures contract work. Most importantly, there must be active market participation. A successful energy futures contract needs 10,000 to 20,000 lots of open interest. If that is not attained, the contract will most likely fail or will need more time to gain industry acceptance. The other important feature about open interest is that exchange-traded options contracts can only be

³ Pennsylvania Jersey Maryland Interconnection.

⁴ New York Independent System Operator.

launched when we have enough liquidity in open interest in the futures market so that it can trade successfully.

I.C.8.3.3 Options on Energy Futures

As energy futures contracts become more liquid, exchanges usually launch options contracts. Exchange-traded options contracts are usually launched after 10,000 to 20,000 lots of open interest are developed in the futures contract. Options on energy futures are traded on the same exchanges that trade the underlying futures contracts and are standardised with respect to the quantity of the underlying futures contracts, expiration date, and strike price (the price at which the underlying futures contract can be bought or sold). As with futures, exchange-traded options positions can be closed out by offset, which is the execution of a trade of equal size on the other side of the market from the transaction that originated the position. Options trading models are very well developed for energy trading in oil and gas futures.

On NYMEX, for example, there is a wide range of option contracts. Aside from the traditional European and American style options, they offer:

- average price options (also called Asian or average rate options). These are settled against the average of prices for an underlying commodity for a specified period. As explained in Chapter I.B.9, the averaging process reduces volatility and hence the option premium.
- calendar spread options. This contract is on the price differential between two delivery dates for the same commodity. It helps market participants manage the risk of changes in the price spread.
- crack spread options. The crack spread is the difference between the price of crude and refined products. This contract helps refiners and other market participants efficiently manage the risk of changes in this differential.

Example I.C.8.2: Contract specifications – Light, sweet crude oil options

Trading Unit: One NYMEX Division light, sweet crude oil futures contract.

Price Quotation: US dollars and cents per barrel.

Trading Hours (all times are New York times): Open outcry trading is conducted from 10:00 a.m. until 2:30 p.m.

Trading Months: Thirty consecutive months, plus long-dated options at 36, 48, 60, 72 and 84 months out on a June/December cycle.

Minimum Price Fluctuation: \$0.01 (1¢) per barrel (\$10.00 per contract).

Maximum Daily Price Fluctuation: No price limits.

Last Trading Day: Trading ends three business days before the underlying futures contract.

Exercise of Options: By a clearing member to the Exchange clearing house no later than 5:30 p.m. or 45 minutes after the underlying futures settlement price is posted, whichever is later, on any day up to and including the options expiration.

Strike Prices: Twenty strike prices in increments of 50¢ per barrel above and below the at-the-money strike price, and the next 10 strike prices in increments of \$2.50 above the highest and below the lowest existing strike prices for a total of at least 61 strike prices. The at-the-money strike price is nearest to the previous day's close of the underlying futures contract. Strike price boundaries are adjusted according to the futures price movements.

Margin Requirements: Margins are required for open short options positions. The margin requirement for an options purchaser will never exceed the premium.

Trading Symbol: LO

Source: New York Mercantile Exchange

I.C.8.3.4 Hedging in Energy Futures Markets

Energy risk management does not eliminate risk. It only shifts it. Hedging is a strategy for price risk to be shifted by using financial instruments such as futures contracts, price swaps or options to shift risk between buyers and sellers. The hedger takes a position in the financial market (futures or OTC) that is equal and opposite to the position that exposes price risk in the cash market, and locks in prices, costs and profit margins. Physical delivery is not anticipated.

The hedger uses risk management tools such as futures and price swaps to protect a physical position or other financial exposure in the market from adverse price movements that would reduce the value of the position. The purpose of a hedge is to avoid the risk of adverse market movements resulting in major losses in the physical market. Because the physical cash markets and futures markets do not always have a perfect price correlation relationship, there is no such thing as a perfect hedge, so there is almost always some profit or loss. In futures markets, hedging involves taking a futures position opposite to that of a cash market position. The seller of the commodity seeks protection against downside price moves and a buyer seeks price protection against upside price moves. For example, an oil producer would sell crude oil futures against its production. Thus, an energy producer or consumer might look at buying or selling energy futures against their price risk exposure in anticipation of a market price increase/decrease.

The hedge position is established to buffer against day-to-day market fluctuations. It is a form of insurance in that the hedger pays an upfront cost to ensure price certainty. The benefits of active energy risk management using futures contracts are to stabilise cash flows, more closely match balance sheet assets and liabilities, reduce transaction costs, decrease costs of storage, lock in 'cost of carry' forward profits, and to minimise the capital at risk needed to carry inventories.

Example I.C.8.3: Electricity producer fears a price decline

In this example, an independent power production company is at risk that falling prices will reduce profitability. It stabilises cash flow by instituting a managed short hedging strategy on the electricity futures market.

On 1 February, the bulk power sales manager at a southeastern utility projects that he will have excess generation for the second quarter and notices attractive prices in the futures market for the April, May and June contracts. The manager arranges to deliver this excess power at the prevailing market price in April, May and June. However, he wants to capture the market prices now, rather than be exposed to the risk of lower prices in the spot markets. The action the utility takes to protect the company from this risk is to sell Entergy electricity futures contracts for those months.

In the futures market, the producer sells 10 futures contracts for each of three months, April, May and June at \$23 per megawatt-hour (MWh), \$23.50 and \$24, respectively. Assuming a perfect hedge, the futures sales realise \$169,280 for the April contracts (10 contracts \times 736 MWh per contract \times \$23 per MWh = \$169,280), \$172,960 for May contracts (10 \times 736 \times \$23.50) and \$176,640 for June contracts (10 \times 736 \times \$24), for a total of \$518,880.

On 29 March, the utility arranges to deliver 7360 MWh of April pre-scheduled power in the cash market, the equivalent of 10 contracts, at the current price which has fallen to \$22 per MWh, and receives \$161,920. That is \$7360 less than budgeted when prices were anticipated at \$23 per MWh.

Simultaneously, the producer buys back the April futures contracts to offset the obligations in the futures market. This also relieves it of the delivery obligation through the Exchange. The April contracts, originally sold for \$23 (\$169,280), are now valued at \$22 per MWh, or \$161,920. This yields a gain in the futures market of \$7360. Therefore:

the cash market sale of	\$161,920 (7360 \times \$22/MWh)
plus a futures gain of	\$ 7360
equals a net amount of	\$169,280, or \$23 per MWh, the budgeted sum for April.

As cash prices continue to be soft for the second quarter, the hedge looks like this:

	Cash Market	Futures Market
Feb. 1		Sells 10 Entergy electricity contracts in each of April, May, June for \$23, \$23.50, \$24, respectively
Mar. 27	Sells 7360 MWh at \$22	Buys back 10 April contracts, \$22
Apr. 26	Sells 7360 MWh at \$23	Buys back 10 May contracts, \$23
May 26	Sells 7360 MWh at \$23.25	Buys back 10 June at \$23

Financial result	April	May	June	Quarter
Expected Revenue	\$169,280	\$172,960	\$176,640	\$518,880
Cash market sales revenue	\$161,920	\$169,280	\$171,120	\$502,320
Futures market gain (loss)	\$7,360	\$3,680	\$5,520	\$16,560
Actual revenue	\$169,280	\$172,960	\$176,640	\$518,880
				\$23.50 per MWh

What happens to the power production company's hedge if prices rise instead of fall?

In that case, assume the cash market rises to \$24, \$24.50, and \$25. The power producer realises \$176,640 on the cash sale of 7360 MWh for April, but sold futures at \$23 in February, and now must buy them back at the higher price, \$24, if it does not want to stand for delivery through the Exchange. The 10 contracts are valued at \$176,640 which is what the company must pay to buy them back, incurring a \$7360 loss on the futures transaction. Therefore:

the cash market sale of: \$176,640 (7360 × \$24/MWh)
 minus a futures loss of: \$7360
 equals a net amount of: \$169,280, or \$23 per MWh, the budgeted sum for
 April.

As cash prices continue to be firm for the second quarter, the hedge looks like this:

	Cash Market	Futures Market
Feb. 1		Sells 10 Entergy electricity contracts in each of April, May, June for \$23, \$23.50, \$24, respectively
Mar. 27	Sells 7360 MWh at \$24	Buys back 10 April contracts, \$24
Apr. 26	Sells 7360 MWh at \$24.50	Buys back 10 May contracts, \$24.50
May 26	Sells 7360 MWh at \$25	Buys back 10 June at \$25

Financial result	April	May	June	Quarter
Expected revenue	\$169,280	\$172,960	\$176,640	\$518,880
Cash market sales revenue	\$176,640	\$180,320	\$184,000	\$540,960
Futures market gain (loss)	(\$7,360)	(\$7,360)	(\$7,360)	(\$22,080)
Actual revenue	\$169,280	\$172,960	\$176,640	\$518,880
				\$23.50 per MWh

The average price of \$23.50 per MWh represents an opportunity cost of \$1 per MWh because cash market prices averaged \$24.50 during the period of the hedge. The producer is comfortable with this because it is within the tolerance for risk that the risk management committee set at the time the positions were opened. Managing a hedge strategy is an evolving process. While hedges serve to stabilise prices, risk management targets can be re-evaluated in future periods as market and financial circumstances change.

Source: New York Mercantile Exchange, *A Guide to Energy Hedging*

Example I.C.8.4: Petroleum marketer’s long hedge, rising and falling markets

On 7 September, the New York Harbor price for heating oil is 55¢ and the cash market price at the fuel dealer’s location is 54¢ a gallon, a 1¢ differential, or basis, between New York Harbor and the retailer’s location. The dealer agrees to deliver 168,000 gallons to a commercial customer in December at 70¢ per gallon. On 7 September, he buys four December heating oil contracts (42,000 gallons each) at 57¢, the price quoted that day on the Exchange’s NYMEX Division, at a total cost of $42,000 \times 4 \times \$0.57 = \$95,760$.

Case 1. Rising prices

On 25 November, the fuel dealer buys 168,000 gallons in the cash market at the prevailing price of 59¢ a gallon, a 1¢ differential to the New York Harbor cash quotation of 60¢, at a cost of \$99,120. He sells his four December futures contracts (initially purchased for 57¢) at 60¢ a gallon, the current price on the Exchange, realising \$100,800 on the sale, for a futures market profit of \$5040 (3¢ a gallon). His cash margin is 11¢ (the difference between his agreed-upon sales price of 70¢ and his cash market acquisition cost of 59¢ for a total of \$18,480 (\$0.11 per gallon × 168,000 gallons).

	Cash market	Futures market
7 Sept.		Buys four December futures contracts for 57¢ per gallon
25 Nov.	Buys 168,000 gallons	Sells four December heating oil futures at 59¢ per gallon for 60¢ per gallon

We have:

a cash margin of \$18,480 or 11¢/gallon
 plus a futures profit of \$5,040 or 3¢/gallon
 equals a total margin of \$23,520 or 14¢/gallon

Case 2. Falling prices

On 25 November, the dealer buys 168,000 gallons at his local truck loading rack for 49¢ a gallon, the prevailing price on that day, based on the New York Harbor cash quotation of 50¢ a gallon. He sells his four December futures contracts for 50¢ a gallon, the futures price that day, realising \$84,000 on the sale, and experiencing a futures loss of \$11,760 (7¢ a gallon).

	Cash Market	Futures Market
7 Sept.		Buys four December heating oil futures at 57¢ per gallon
25 Nov.	Buys 168,000 gallons	Sells four December heating oil futures for 49¢ per gallon for 50¢ per gallon

We now have:

a cash margin of	\$35,280
minus a futures loss of	\$11,760 (7¢/gallon)
equals a total margin of	\$23,520 (14¢/gallon)

In summary, the fuel retailer guarantees himself a margin of 14¢ a gallon regardless of price moves upwards or down in the market. With the differential between cash and futures stable, as in cases 1 and 2, spot-price changes in either direction are the same for both New York and the marketer's location. As a result, a decline in the futures price, which causes a loss in the futures market, is offset cent-for-cent by the increase in the cash margin.

Source: New York Mercantile Exchange, *A Guide to Energy Hedging*

Speculators, on the other hand, are usually not active in the physical markets as either producers or consumers of the physical commodity. They take no physical commodity position. Their risk appetite is significantly higher than that of the hedger. What attracts the speculator is the potential for profit. In effect, the speculator assumes the hedger's risk and adds liquidity to the market. Markets would die without the active participation of speculators since their view on the market is important for market liquidity.

I.C.8.3.5 Physical Delivery

The energy futures markets have an active underlying physical commodity market. Because of this factor, some futures contracts go to contract expiry for physical delivery. It should be remembered that futures contracts should not be used to gain physical supply but only to manage the price risk of that supply. Although generally no more than 2% of energy futures contracts ever go to physical delivery via the exchange, the NYMEX and IPE offer several methods for physical delivery of their expiring futures contracts. The NYMEX physical delivery point for its crude oil contract is Cushing, Oklahoma, for its heating oil and gasoline contracts is New York Harbor, and for natural gas contract is Erath, Louisiana. The IPE's gasoil contract was launched in 1981 and is delivered into the barge market for the Antwerp, Rotterdam and Amsterdam (ARA) region (barge lots are typically 50,000 to 100,000 tonnes in the physical market). The IPE Brent crude oil contract is the only large energy futures contract that does not go to physical delivery on expiry, going instead to cash settlement.

Companies which do choose to deliver, however, have several options. They can choose standard delivery (as per the specification of the futures contracts laid down in the rule book of the futures exchange) or they can attempt to arrange an ADP. This would normally happen if someone wishes to deliver some energy which is not in keeping with the specifications of the futures

contract (so it cannot be delivered via the exchange/clearing house procedure) or two parties are stuck with a position and they just wish to negotiate directly with one another on some cash settlement. In ADP transactions, market participants release both the exchange and their clearing broker member from all liabilities related to the delivery negotiated between parties. Traders and hedgers can also execute an EFP.

Companies using energy futures contracts for hedging purposes are often not interested in making or taking delivery at the specified locations. More often than not, a hedger using futures finds it more economical to make or take delivery of physical energy elsewhere, under terms that differ from those of the futures contract. An EFP provides the mechanism for such transactions and is usually the preferred method of delivery because it provides greater flexibility. EFPs allow companies to choose their trading partners, delivery site, the grade of product to be delivered, and the timing of delivery. The EFP mechanism allows buyers and sellers to execute their physical energy market transaction on the basis of negotiated price. After both parties to an EFP agree to such a transaction, the price at which the EFP is to be cleared is submitted to their futures broker who in turn submits it to the futures exchange which then registers the trade. The price of the futures position created by the EFP can be outside the daily trading range of that futures market. This is the nominal price of the EFP. The EFP parties can then effect the actual physical exchange at a price they negotiate between themselves.

Example I.C.8.5: EFP to initiate a position

On 7 August, an oil refiner who wishes to protect a portion of his products inventory wants to sell to protect against falling prices. At the same time, a diesel fuel distributor is concerned about rising prices and looks to buy to protect his forward purchases.

They agree to a price of diesel fuel, net the basis, and register the EFP with the Exchange. Once registered, both parties have instituted futures positions at a price which reflects the exact basis between NYMEX Division heating oil futures and the regional rack price for diesel fuel.

On 16 September, the diesel refiner arranges with the distributor for the physical delivery of the fuel. At that time, the refiner and the distributor independently offset their futures positions on the Exchange.

In this example, the long and short hedgers have ‘swapped’ futures obligations (thus terminating their contract obligations on the Exchange before their futures contracts mature) in consideration of their exchange of physical market positions. The transaction occurs at the price, location and time negotiated by the parties.

In order to engage in an EFP, it is not necessary for both sides to be in the futures market when the EFP is initiated except during EFP-only sessions. A futures market long, for example, might effect an EFP with a cash market long. The net result of the transaction would be delivery of a commodity to the futures hedger and the assumption by the physical market participant of the hedger's futures market contracts.

Source: New York Mercantile Exchange, *A Guide to Energy Hedging*

I.C.8.3.6 Market Changes: Backwardation and Contango

The shape of the forward price curve has been explored generally in Chapters I.A.7 and I.B.3. Contango is a market condition where prices are progressively higher in future contract months; this is considered to be the 'normal' condition of markets since cost of carry (funding, storage and the like) is generally positive. Backwardation is a market condition where prices are progressively lower in future contract months. Thus, when the price for a contract month nearer to the present time is higher than the price for a contract further into the future, the market is said to be in backwardation. As explained in Section I.C.7.3.1, this typically occurs when prices are high because supplies are tight. When markets are in backwardation the strike price for a calendar spread options contract will be positive. When the price curve is in contango, strike prices of calendar spread options contracts will be negative. A negative price is not unusual in spread relationships.

In contango markets, the producer, who is a seller of oil, would seek downside protection by buying puts; an oil buyer would purchase calls. A crude oil producer with excess storage capacity can make money when the price curve is in contango by purchasing the cheaper prompt month and selling the more expensive deferred contract month.

When the markets are in backwardation, however, spare storage capacity is an asset that generates no cash flow. Selling put options on calendar spreads generates cash flow, and having the asset as a backstop enables the oil company to sell the put. Additionally, in a steeply backwardated market, it can be costly to buy back a hedge after it has appreciated in value on its way to becoming the prompt month. Buying calls on the calendar spread can reduce such costs, and can complement the short hedge by allowing for participation in the rising market.

I.C.8.4 OTC Energy Derivative Markets

Nearly all the key terms of an OTC derivatives deal are negotiable, which means that the pricing reference, the payment terms and the volume can all be adjusted to suit the counterparties to the deal. In effect, they are customised transactions. Fortunately, for risk management purposes, the core energy markets, like the larger oil, gas and electricity (power) markets, have some active and

fairly standardised OTC contracts. They are standard both in their floating price reference, and in the sort of minimum contract volume that would normally be traded.

Brief explanations of some of the most common transaction types are given below, along with brief illustrations of their application in energy markets:

Forward Contracts

A contract to buy/sell with future delivery. The contract sets the price (or price formula) in advance. For example, an oil refiner may purchase crude oil in the forward market to hedge against possible price rises. See Chapter I.B.3 for a general discussion of forward contracts.

Option Contracts

A contract giving the buyer of the contract the right to buy or sell at a pre-specified price. For example, a gas supplier might buy a put option giving the right to sell gas at a pre-specified future price. In addition, option contracts can be based on the crack spread (the difference between the price of refined products and crude) or on the calendar spread (the difference between prices for different maturities). See Chapter I.B.5 for a general discussion of option contracts.

Swaps

A swap contract provides for the parties to exchange a series of cash flows generated by underlying assets. See Chapter I.B.4 for a general discussion of swap contracts. There is no transfer of any assets or principal amounts. For example, in a simple crude oil swap a refiner and oil producer enter a five-year swap with monthly payments. The refiner pays the producer a fixed price per barrel. The producer pays a variable price which could, for example, be based on the settlement price of a futures contract on the final trading day of the month. The notional amount of the contract is 10,000 barrels. Payments are netted so that a payment of differences only occurs.

Consider another example where a large consumer of electricity purchases electricity from a local distribution company at variable market prices. He wishes to hedge against increases in electricity prices and can do so by entering a fixed for floating swap contract with monthly payments. The notional principal for the contract (the swap contract load) is 10 MWh and the swap fixed price is \$50 per MWh. If the average spot market price exceeds (is less than) the agreed strike price, then the counterparty (consumer) will pay the consumer (counterparty) the difference. Suppose that the market clearing price for the month is \$60 per MWh, then the counterparty pays the consumer \$100, being the swap contract load of 10 MWh multiplied by the price difference (\$60 – \$50 per MWh).

Basis Contracts

A basis contract helps to hedge locational and product differences between exchange-traded standard contracts and the actual exposure of the user. Suppose that a local distribution company needs to purchase gas and decides to hedge in the futures market. The futures contract allows the company to lock in a future Henry Hub price. The actual price paid for gas in the local market may, however, differ from the Henry Hub price. If the local price increases by more than the Henry Hub price, then the company will incur a loss due to basis risk. A basis swap is an OTC contract that fixes the price gap to give the company complete price certainty.

Today regulated futures exchanges have been slow to react to changing markets and unable to launch many successful energy futures contracts due to the success of the OTC energy markets. The effectiveness of the OTC energy markets is most clearly seen in Asia, which overtook Europe as the second largest oil-consuming region in the world several years ago. However, Asia still does not have a liquid and internationally recognised futures exchange for energy market. This is because its needs for energy-related derivatives contracts seem to be well served by the established OTC market, for which Singapore is the key trading hub.

In Asia, the vast majority of physical transactions and OTC swaps are priced using the industry-recognised publication Platts, which is a division of McGraw-Hill. Platts publishes a daily assessment of the price of any given crude oil or oil product in any given location, and also publishes an assessment of the forward curve. These daily value assessments are based on the aggregated bids and offers from many brokers and dealers around the world during a specified time window for each geographic region – usually towards the end of each business day in each major time zone: Asia (Singapore), Europe (London), USA (New York, and then the West Coast).

Price swaps are usually priced off the monthly average of these Platts assessments and lead to a monthly financial payment equivalent to the difference between the traded fixed price and the calculated average floating price multiplied by the contractual monthly quantity. Only the difference is paid and there is no exchange of physical energy, hence no delivery risk.

The futures exchanges have reacted to the OTC markets which tend to be longer-term in nature by launching clearing house platforms. By clearing these OTC contracts on an established futures exchange which is government-regulated, the contracts become quasi-futures contracts. OTC clearing for NYMEX through Clearport and the IPE through the London Clearing House have created more liquidity for the exchanges. The structure of the energy derivatives markets has been that futures contracts are very liquid and traded in the front months and the deeper OTC

markets are used for longer-term hedging needs. By linking OTC to exchange-traded through clearing, the markets now show many more points of trade and eliminate performance risk since the exchange clears for both buyer and seller, and assumes counterparty risk. Exchange members guarantee performance and are overseen by government regulatory agencies. The effect is that the core OTC energy derivatives are becoming more and more indistinguishable from futures trades, and there has been increasing convergence between OTC and futures contracts.

I.C.8.4.1 The Singapore Market

The Singapore market is the trading hub of Asia and is oriented to cargo size shipments, so individual transactions, sometimes referred to as ‘clips’, are quite large compared to an IPE Brent futures contract which is a minimum trade of 1000 barrels. Almost everything in Singapore market is sold in 50,000 barrel clips and the typical cargo size is 150,000 barrels; high sulphur fuel oil used for ships bunkering is the only exception. To put this into perspective in terms of growth, in 1998 the Singapore swaps market was estimated to be around 150 million barrels per month. There is an active OTC swaps broking community in Singapore (there are around 10 active OTC broking companies there) which adds market liquidity by assisting price discovery in the market and by developing two-way markets for buyers and sellers of oil. Typically most Asian oil products and related crude oils can be traded up to 18 months forward, with most of the liquidity in 1–12 months forward markets. Beyond 2 years forward, the number of participants quoting prices become more limited (mainly to large bank traders and major international oil companies). The key products traded are:

- Singapore gasoil 0.5%
- Singapore jet fuel (kero)
- Regrade – the spread between Singapore gasoil and Singapore jet fuel
- 180 CST fuel oil
- 380 CST fuel oil
- Naphtha
- Tapis crude oil – Malaysian exported crude oil
- Dubai/Oman crude oil – Middle East market crude, meaning many Asian refiners are buying crude oil as feedstocks for their refineries on a Dubai/Oman pricing basis

I.C.8.4.2 The European Energy Markets

Europe has a very active and well-developed OTC oil market. Unless otherwise specified by traders or brokers, a quote on European-based swaps will normally price against the mean average of Platts high/low assessment of the relevant European physical market. Fixed for floating swaps and caps and collar options will normally be available in the following markets:

- Premium unleaded barge f.o.b. ARA

- Premium unleaded crack swap North West Europe (NWE)
- Naphtha c.i.f. NWE
- Naphtha crack NWE
- Jet diff c.i.f. NWE
- Gasoil crack swap NWE
- Gasoil 0.2% cargo c.i.f. NWE
- Gasoil 0.2% barge f.o.b. ARA
- EN590 cargo c.i.f. NWE
- Gasoil 0.2% cargo f.o.b. Mediterranean (MED)
- Fuel Oil 1% cargo f.o.b. NWE
- Fuel Oil 3.5% Barge f.o.b. ARA
- Fuel Oil 3.5% cargo f.o.b. MED
- LPG propane c.i.f. ARA large
- Brent/Dubai swaps
- West Texas Intermediate (WTI) crude oil versus Brent crude swaps

More exotic one-off derivative structures are normally available, given the higher number of participants in the European oil swaps market and the higher liquidity in the plain vanilla market.

Key OTC products include:

- Rotterdam gasoil 0.2% sulphur barges
- Jet fuel, NWE cargoes c.i.f. basis
- Jet fuel, Rotterdam barges f.o.b.
- Gasoil IPE futures look-alike swap
- Fuel oil 1% NWE cargoes c.i.f. basis
- Fuel oil 1% NWE cargoes f.o.b. basis
- Fuel oil 3.5% Rotterdam barges f.o.b. basis
- Fuel oil MED 3.5% cargoes f.o.b. basis
- Dated Brent related swap (dated Brent is spot North Sea oil)
- Brent IPE futures look-alike swap
- Brent bullet swap
- Dubai crude oil swap trade out of London (as well as Singapore)
- EN590 grade gasoil NWE cargoes c.i.f. basis
- EN590 grade gasoil MED cargoes
- Gasoline – Rotterdam Eurograde barges
- Naphtha NWE c.i.f. cargo swap

- European natural gas firm physical, fixed price
- European natural gas firm physical, fixed price, spread Zeebrugge versus National Balancing Point
- UK National Balancing Point indexed OTC Swaps basis NBP97 contract
- LPG Mid East/North Africa/Asia – Saudi CP pricing used as index for OTC swaps
- LNG with Crude related pricing formula - proxy hedging in crude futures/related OTC derivatives markets

I.C.8.4.3 The North American Markets

The most developed and liquid OTC energy derivatives market is in North America. Oil futures trading began in 1978 and the price swap was invented by Chase in 1986. The OTC price swaps markets have not been regulated since the July 1989 regulatory ruling by the US Commodity Futures Trading Commission that the commission would not regulate commodity swaps. This action has been reinforced by the Commodity Futures Modernization Act of 2000.

Price swaps are active for crude oil and petroleum products, natural gas, electricity, coal and emissions. There are very sophisticated OTC options products offered as well. The demise of Enron and merchant energy did not affect the oil markets as much as the gas and power markets in North America since Enron and other merchant power market makers were substantial financial players in gas and electricity trading on both the futures exchanges and the larger OTC markets. They were not substantial players in world oil markets. Investment banks and hedge funds are increasingly stepping in to make up for the loss of trading liquidity in natural gas and power.

Unless otherwise specified by traders or brokers, a quote on a USA-based price swap will normally price against the mean average of Platts high/low assessment of the relevant American physical market. Fixed for floating swaps and caps and collar options will normally be available in the following markets:

- Nymex light sweet crude oil related 1st line futures look-alike swap
- Nymex light sweet crude oil related bullet swap
- Nymex heating oil futures related bullet swap
- Nymex heating oil futures related 1st line futures look-alike swap
- Nymex gasoline futures related 1st line futures look-alike swap
- Nymex gasoline related bullet swap
- New York Harbor #2 heating oil barges
- 1% fuel oil New York Harbor c.i.f. cargo basis

- 3% US Gulf Coast cargo
- US Gulf Coast gasoline 87
- New York Harbor reformulated RFG gasoline 87 barges
- New York Harbor gasoline 87 barges
- Canadian natural gas firm physical, fixed price
- Canadian natural gas firm physical, Canadian gas price reporter
- Natural gas, fixed for float (inside FERC)
- Natural gas, fixed for float (NGI)

The density of the North American OTC energy derivatives market is illustrated by the fact that there have been over 500 different locations for trading for natural gas and electricity on the North American continent. The oil markets are even more complex as there are hundreds of grades of gasoline, for example. The point is that the OTC markets have customised contracts to manage energy price risk for many more commodities than are listed on exchanges. The exchange response has been to clear OTC contracts through their clearing houses. These efforts have been quite successful since their launch in 2002.

I.C.8.5 Emerging Energy Commodity Markets

There are several new emerging markets in the energy world. These include coal, weather derivatives and environmental financial or ‘green trading’ markets.

I.C.8.5.1 Coal Trading

Most electric power in the world is still coal-fired, but risk management for coal is still beginning. The reason is that most coal was formerly priced on long-term contracts. The development of a spot market for coal in recent years has been a slow process but the recent active buying by China, and US electricity deregulation, have brought it into global commodity markets with greater price volatility. China’s position and recent structural changes in its production profile (lots of small mines closures etc.) have been the key to global supply and demand. The USA has more high-quality coal than any other country, with nearly 30% of the world’s bituminous and anthracite coal reserves and 250 years of supply. Australia, Indonesia and Columbia are other high-quality coal exporters. Only China produces more bituminous coal than the USA, but almost all of its production is consumed domestically. US coal exports, chiefly central Appalachian bituminous, make up 16% of the world export market and are an important factor in world coal prices. The importance of coal can be seen from the fact that coal-fired power stations still account for approximately 55% of total electricity output in the USA. Coal is transported by rail or barge. Railroads carry more than half of the coal mined in the USA, often

hauling the coal in unit trains. The US inland waterway system is the other major mode for coal transportation, especially along the Ohio and Mississippi rivers.

The impact on the environment of coal use is a serious issue. Any effort to curtail atmospheric emissions can be expected to involve reduced coal use, even though the amount of air pollution produced by coal burning has been greatly diminished during the past 30 years due to air quality considerations. Therefore, more stringent environmental regulations have created many new arbitrage opportunities between coal and emissions.

Coal is still mainly traded in the OTC market, however, for both the US coal mining and electric power industries, although a NYMEX coal futures contract exists. The NYMEX futures contract specifications have effectively been co-opted by the OTC markets. Today, we are in the midst of a major market entry by many financial houses, and commodity traders such as Cargill, EdF, Morgan Stanley, and Goldman Sachs have entered coal trading markets and are making markets in coal derivatives. There are also very well-established OTC brokers such as TFS, GFI, Spectron, Evolution Markets, and Natsource. For the international coal trading or consuming market, the OTC market dominates coal pricing and hedging.

Coal producers can sell future production contracts to lock in a specific sales price for a specific volume of the coal they intend to produce in coming months. Electric utilities can buy coal futures to hedge against rising prices for their baseload fuel. Power marketers, who have exposure on both the generating and delivery sides of the electricity market, can hedge with coal futures to mitigate their generation price risk, and hedge with electricity futures to control their delivery price risk. Non-utility industrial coal users, such as steel mills, can use futures to lock in their own coal supply costs. International coal trading companies can use futures to hedge their export or import prices. Power generating companies that use both coal and natural gas to produce electricity can use coal futures in conjunction with the NYMEX Henry Hub natural gas futures to offset seasonal cost variations and to take advantage of the ‘spark spread’ – the differential between the cost of the two fuels and the relative value of the electricity generated by each of the two fuels.

For the coal derivatives market to develop, it is important that there is a solid critical base of consumers, producers and financial traders. The good news is that new large entrants on both the consumer and trading sides have entered the market. Also, more financial traders have been looking to join the market. The key is that pricing indices are needed for coal derivatives trading due to the lack of viable futures contracts. In Europe, it is the traditional tonnage delivered into

the ARA region. While the majority of the coal delivered into ARA is shipped from South Africa, it also includes tonnage (depending on market conditions) from the USA and Columbia.

Price indexes are widely used in energy derivatives trading and typically started with price reporters for trade publications. Later brokers started offering their quotes based on their deal flow and now many price quotes are on the Internet. The key is that for emerging markets price indices provide price transparency which stimulates more trading. It began in oil trading and now is used for gas, power and coal. In coal trading, there have been index deals up to 4 years forward. The majority of participants actively trade up to 2 years forward in the cash-settled coal swaps (based on ISDA agreements). The standard documentation used in energy market settlement for the OTC markets is provided by ISDA, the International Swaps Dealers Association.

I.C.8.5.2 Weather Derivatives

Weather derivatives have always had more hype than success as a fungible commodity. Since its start in commodity markets in 1997, there have been over 10,000 weather derivative trades with a notional value of \$15.8 billion, according to the Weather Risk Management Association (WRMA, see www.wrma.org). There are illiquid weather futures contracts on the Chicago Mercantile Exchange, but basically this continues to be a one-off market with no trading facility. Despite all the publicity, it remains a reinsurance financial product and will continue to be situated there, not as a commodity trading market. Features of commodity markets require replication and simplicity of trading in both standardised futures contracts and in OTC markets. Weather derivatives are complex financial structures meeting the customised weather profiles of utility, industrial and agricultural customers. This specialised customisation makes it very difficult to replicate trades and create fungible trading commodities. The reinsurance industry created crop reinsurance products for farmers based on weather over two decades ago, and it seems more than likely that the reinsurance industry will continue to provide the financial cover for the weather derivatives markets. The paucity of weather futures trades and derivatives in general points to that conclusion despite all the publicity about weather derivatives in the trade and financial press.

Weather risk embraces a wide variety of natural phenomena, but if the focus is limited to temperature risk, the marketplace starts to look almost as standardised as any commodity futures market. In fact, the 2002 survey conducted for the WRMA shows that weather derivatives referenced against temperature accounted for over 80% of the total volume. (i.e. heating degree days (HDD)/cooling degree days (CDD)). Temperature-related weather derivatives help people hedge or trade the temperature at certain agreed geographical points around the world. Demand for weather risk derivatives indexed against temperature references is most probably driven by

energy producers and users. The link between the two markets arises since energy demand varies with fluctuations in temperature. The most prevalent use of weather derivatives has been to hedge uncertainty in volumetric demand for energy, due to temperature fluctuations. For this purpose, deals are often referenced to the number of heating (or cooling) degree days in a period. This measures the daily deviation from a reference temperature (e.g. 18°C), and sums the negative (for HDDs) or positive (CDDs) deviations over the period. Weather derivatives are really a reinsurance product, and that is where much of the business is now heading. It is thought in some quarters that climate change risk may add some liquidity for weather derivatives trading.

I.C.8.5.3 Green Trading

The Kyoto Protocol of 1997 called for the trading of emission reduction as a way of levelling out the cost of reducing greenhouse gases. The most important emission for green trading is carbon dioxide, but other emissions include methane and nitrous oxide. Emission reduction credits (of which carbon credits are the most common) can be earned in several ways: upgrading power-generating equipment so that there are fewer emissions per kilowatt-hour generated, planting trees or reducing soil erosion so that more carbon is sequestered or absorbed, and producing electricity using renewable sources such as solar panels and wind turbines. Once carbon credits are earned, they can be sold to other companies. The credit is subtracted from the recipient party's total output of emissions, so that the recipient can meet a voluntary or legislative emissions target. So carbon credits generated in Germany could be used to meet emissions targets in Australia. Some countries/industries are more efficient at producing emission reduction credits and they can exploit their comparative advantage through green trading. Since greenhouse gases released into the atmosphere can wander globally, the source of the emission credits is not important.

Green trading encompasses the convergence of the capital markets and the environment. It is the first global market that we have seen since crude oil trading and presents many opportunities for both the energy industry and financial institutions. The energy industry, the world's leading emissions polluter, will be the leading supplier of environmental solutions because it is good business. Today, the industry is at a turning point on global warming as carbon intensity continues to grow whilst time to stabilise carbon dioxide and other greenhouse gas emissions is limited. The carbon emissions footprint of the major oil companies is evidenced by their global oil and gas production, refining and transportation, and their involvement in the power industry continues to expand. Environmental issues are becoming corporate financial issues. Greater financial disclosure of corporate environmental risks, including climate change, has raised the issue of environment as a corporate fiduciary responsibility. Increasingly, environmental and

financial performance of companies is intertwined, and these new financial risks and liabilities will prompt change and market creation.

Environmental financial products for sulphur dioxide and nitrous oxides have been successful in controlling US pollution since 1995 and 1999, respectively. A \$6 billion environmental market today pales into insignificance compared to a \$2 trillion energy derivatives market, but the growth trajectory suggests that green trading markets today should be compared with the oil markets of 1978. However, this time, maturation will be global and simultaneous as carbon trading regimes take root in the EU, Asia, Australia and North America. Thus far, only a couple of hundred trades have taken place, but estimates suggest that a \$3 trillion commodity market may emerge over the next 20 years. The dollar value of this market is enticing, but the reality is that the global energy industry will be the primary supplier of liquidity to this market followed by the agricultural industry. Both industries are active in commodity trading. Green Trading includes trading not only in greenhouse gas emissions but also in renewable energy and the financial value of energy efficiency. Cross-commodity arbitrage opportunities are self-evident as oil, gas, coal and power, like weather derivatives, have environmental dimensions.

There will be two stages in the development of the international carbon market. Now, in stage one, carbon reduction credits are being created.⁵ Trading covers many years because, thus far, there has not been an allocation of sufficient units to have a spot market and because these are project-based reductions. Capital is required and forward commitment cannot be banked. If the World Bank is buying a 10-year stream of reductions, a bank loan would usually be available to implement the project. Consequently, there are still these structured trades with very large volumes. Early speculative trading and some hedging of risks are taking place; as is a transformation from the environmental department to the risk manager, in energy companies and energy end-user groups, as some major corporations treat the greenhouse gas issue as a financial issue. In this, the early stages of the market, carbon finance is playing a bigger role and, probably during 2005, a liquid spot market will develop.

The second stage of development of the carbon market will be towards a mature and liquid market and, over the next 10 years, there will be linked markets and then index markets. We shall see spot trading, high volumes, advanced brokerage, similar to the power and gas markets, and a growth in carbon finance.

⁵ A carbon credit gives the right to emit a certain amount of carbon into the atmosphere. Carbon makes up 80% of greenhouse gases.

As almost all environmental financial contracts such as those in sulphur dioxide and carbon dioxide are traded on the OTC markets, there is an opportunity for exchanges such as IPE and NYMEX to offer OTC clearing, which would effectively make them quasi-futures contracts under government oversight. This could help to make them more acceptable to risk managers. The IPE recognised this opportunity in April 2004 and has linked its platform to the Chicago Climate Exchange (www.chicagoclimateex.com) in order to trade emissions in the EU.

I.C.8.5.4 Freight-Rate Swaps

Freight is an integral part of the global energy business, either directly or indirectly, as the cost of shipping oil or gas around the world affects power prices in some way. The biggest and in the past the least hedgeable risk in an international oil transaction, for example, has been the freight-rate movements. Tanker freight swaps are a beneficial risk management tool for the energy industry because, in the past, companies sometimes had unhedgeable freight exposure on both physical and paper positions. The main participants in this market are petroleum products traders and shipping charterers, as well as tanker owners and banks. Trading volume can be tailored to users' needs, and the most commonly traded lot is 10,000 tonnes per contract month. Most frequently talked tenures for these swaps are 2–3 months in the future, while bid/offer quotations are usually available for up to 6 months in the future.

I.C.8.5.5 Derivative Forward Price Curves

A very recent development has been the introduction, by a few specialist firms, of independent assessments of the forward curves for an increasing number of global energy derivatives markets. This enables both bankers and end users to have a trusted third-party forward curve for day-to-day valuation and accounting purposes. The Enron scandal of 2001 rocked many shareholders' confidence in companies' use of energy derivatives as well as their pricing and accounting as Enron 'tilted' the curve (Enron fraudulently reconfigured their forward price curves to meet quarterly financial targets). Therefore, the opportunity to utilise third-party market assessments of forward curves is a very positive step towards ensuring that a reasonable value is attached to derivatives. Indeed, under the new accounting regimes of FAS133 and international accounting standards, derivatives need to be revalued on a regular basis, even if they are employed by an end user such as a power producer or airline.

I.C.8.6 The Future of Energy Trading

I.C.8.6.1 Re-emergence of Speculative Trading?

Enron and the merchant power sector in 2001–3 depended on highly leveraged trading to boost paper profits. Since their demise much of the energy industry has returned to the relative safety of trading around assets and marketing activities, avoiding more speculative trading. However,

energy markets have become increasingly characterised by increasing prices and price volatilities across all energy commodities in recent years. That price volatility has been very attractive for speculator traders at investment banks and now hedge funds.. Oil markets are now booming and were not at all affected by the Enron collapse and, as a result of geopolitical issues, the relative weakness of the US dollar, and other supply/demand factors, these higher prices are sustainable with increased price volatilities set to be the norm. The future for North American natural gas is similar as supply and production declines have also resulted in higher sustainable prices and increased price volatilities. These sustainable higher oil and gas prices are evidenced by the lack of investment in energy infrastructure in both the upstream and downstream segments of the industry, the higher costs of finding and developing new energy resources, no conservation effect due to these higher prices and the complacency of energy companies to make substantial profits without taking on new financial risks. Meanwhile, robust demand for coal is also apparent with over 90 new coal plants in line for construction in the USA as the attraction of natural gas as a generation fuel recedes. Electric power is also showing price volatility. It is a combination of this price volatility and available trading talent that is creating the opportunity for hedge funds. With over 200 energy hedge funds already playing or set to play in energy commodities, these funds are primed to bring more risk capital to bear in energy markets. They also bring sophistication, liquidity, the risk culture and trading acumen to bear on energy markets and have access to readily available experienced trading resources that were let go by the mega-merchants. While new hedge funds are being created specifically for the energy trading opportunity, existing larger hedge funds are also planning to enter energy markets.

I.C.8.6.2 Electronic Energy Trading

The electronic energy solution is upon us. The willingness to embrace the more flexible OTC market at the expense of more efficient exchange mechanisms suggests that an open electronic platform combined with the flexibility of the OTC market is the way forward. The trading platform of the future must be able to match identical bids and offers as well as be flexible enough to negotiate deals that cannot easily be matched. While NYMEX has an electronic platform called ACCESS that trades after trading hours on the floor are completed, there is a larger Internet platform that is the chief competitor to NYMEX. The ICE was launched in August 2000 and does not take title or any participatory interest in any transactions on the exchange. In a sense, the exchange provides the arena and defines the rules. The actual playing of the game is left entirely to the users of the exchange. It is a level playing field for all, without favouritism or control by a chosen few, open to any user. Recent statistics show that during peak usage hours the exchange typically has over 3000 users logged on simultaneously.

The exchange lists over 600 unique products covering a variety of commodities, structures, and settlement terms including:

- oil, natural gas, electric power, precious metals, emissions and weather;
- physical delivery and financial cash settlement;
- forwards, swaps, options, spreads, differentials, complex derivatives.

The ICE is primarily a matching system. It allows credit and risk managers from all registered companies to specify and pre-clear credit for trading with each other. This is done using the Counterparty Credit Facility which can open or close credit with each registered user at any time, set tenor limits and set daily dollar limits for trades with each user. In addition to bilateral credit, the ICE supports clearing services for some major products, supported by the London Clearing House.

The ICE brings parallel trading in IPE energy futures for crude oil and gasoil to the Interchange, alongside simultaneously OTC trading on the ICE. Electronic trading sessions run in parallel with open outcry sessions on the IPE London trading floor. As a result, IPE's Brent crude oil contract trades for an uninterrupted 20 hours daily, including pre-market and after-hours trading. Known as 'e-Brent', the electronic oil futures contract makes accessible a critical global energy benchmark to hedgers and speculators across the USA, Europe and Asia.

The growth of electronic energy trading is a slow evolutionary process that has nothing to do with technology. The change that must be facilitated is human behaviour. On both NYMEX and IPE, it will be a slow process of change for electronic trading to migrate from the floor to cyberspace. Floor traders and upstairs brokers still have the client relationship, knowledge of markets, information flow and other factors that cannot be gleaned from screens.

I.C.8.6.3 Trading in Asian Markets

Energy trading and the use of energy risk management tools have been slow to evolve in Asian energy trading. This has been primarily due to the scarcity of natural resources in the region and its focus on security of supply. Asian energy markets are still oriented to security of supply issues over financial risk management and are just beginning their ascendancy into much more mature financial markets; consequently, there are no viable energy futures contracts in Asia as most trade is bilateral and off exchange. Risks in energy markets are pervasive today. This is largely due to deregulation, globalisation, and privatisation trends in many countries, coupled with robust energy demand growth. Borrowing heavily from the institutional memory of well-developed New York and London capital markets, energy trading and risk management are on an upward trajectory in Asia fuelled by growing oil and gas dependencies and the need for more electric power.

While short-term physical oil trading has always existed in most Asian countries, the energy complex is now broadening to include gas, power, petrochemical, coal and weather risk management. Lurking on the horizon is emissions trading to reduce plant emissions and reduce greenhouse gas emissions. Asia is now primed to embrace the active use of energy derivatives and much more sophisticated trading techniques and financial engineering.

Today, Asia is ripe for fundamental change in its trading and risk profile, with the region experiencing rapid economic growth fuelling increased needs for crude oil and refined products supply. The largest consumer in the region, China, is currently the world's second largest oil consumer, behind the USA, and has recently surpassed Japan. Chinese oil consumption will reach some 10.9 million barrels per day by 2025, with net imports of 7.5 million barrels per day, in order to support its domestic growth, giving it a major role in the world oil market. This growth in demand is driving rapidly increasing supply chain complexity as new trading patterns develop. Growth in the region, and in China specifically, is leading to the development of new supply markets in both the Middle East and Russia and in new infrastructure construction from the point of supply to the refinery and beyond. Since most shipments are undertaken by water, the new infrastructure includes tankers, terminals, storage facilities, refineries and overland distribution systems.

While trading remains largely based on term OTC contracts without a standard regional marker for price transparency, it is this supply chain complexity that will drive costs and risk in the medium term, both in China and the Asia Pacific region generally. The risks and costs involved are becoming too great to rely on the in-house developed or spreadsheet-based energy trading. Thus, transaction and risk management systems are used more today by many of the region's major energy companies. The transition in the market from monopoly to competitive markets has fundamentally changed how utilities and others buy and sell electricity. It is now the beginning of the transition to competitive markets and trading in Asia Pacific.

I.C.8.7 Conclusion

Twenty-six years after the first successful oil futures contract we are now seeing the development of a true multicommodity market that encompasses oil, gas, power, coal, freight rates, weather and green trading. Energy commodity trading is evolving into many areas of the energy complex and extending into emerging commodity markets such coal, emissions and weather trading. Convergence (a term often much overused) is actually now upon us as multicommodity arbitrage is the watchword of today's energy trader. Increased price volatility, the extra liquidity provided by financial institutions, and a greater risk appetite are three major factors that make the present time the real dawn of energy trading. Energy risk management has become not only a fiduciary

responsibility but also a core competency of energy companies. Broader penetration into the emerging markets of the developing world, and particularly Asia, shows that there are no barriers to entry in trading on the Internet. The true financialisation of the energy commodity markets is upon us.

The extension of energy commodity trading expertise for natural gas began in North America in the early 1990s and spread to Europe in the late 1990s. A global gas market is now emerging due to the emergence of LNG as a commodity market. Both physical and financial markets for electric power are now established in North America, Europe and Australia, and are beginning in Asia. Since 1995 there has been an active emissions market for sulphur dioxide in North America, and this environmental financial market has proved to be the template for global trading of carbon dioxide. A true financial market emerges for energy trading emerging globally.

While the oil and gas industry has been gradually using energy risk management tools to manage its financial risk over past two decades., the unprecedented price volatility of the past few years in oil, gas and power are actually accelerating industry's adoption of both financial instruments and Internet energy trading. While liquidity on the Internet today still remains low, it is the vehicle for global commodity trading and will be a tool for establishing trading throughout those areas of the world that have no liquid energy commodity contracts, particularly in Asia. The costs are now lower and the security is more reliable. Internet trading will force the migration of the energy exchanges to the web for the paper trading of oil, gas and power. Moreover, the other industry trends of market consolidation, market liberalisation, globalisation and privatisation are creating greater risks to be managed proactively. The medium for that risk management will be the Internet.

Commoditisation has been accelerating due to the financial wherewithal and risk capital of investment banks, multinational energy companies and now the hedge funds. The good news is that commodity markets need more players to provide and develop fungible financial products and provide liquidity. The better news is that more volatile energy commodity markets have brought more risk and there is a great need for more proactive energy risk management than ever before.

II.A Foundations

Keith Parramore and Terry Watsham¹

Our mathematical backgrounds are accumulated from time of birth, sometimes naturally, sometimes forcibly! We cannot hope to cover even a small part of that ground in this introductory section. We can only try to nudge the memory, and perhaps to fill in just a few relevant gaps to get you ready for the sections that follow.

II.A.1 Symbols and Rules

II.A.1.1 Expressions, Functions, Graphs, Equations and Greek

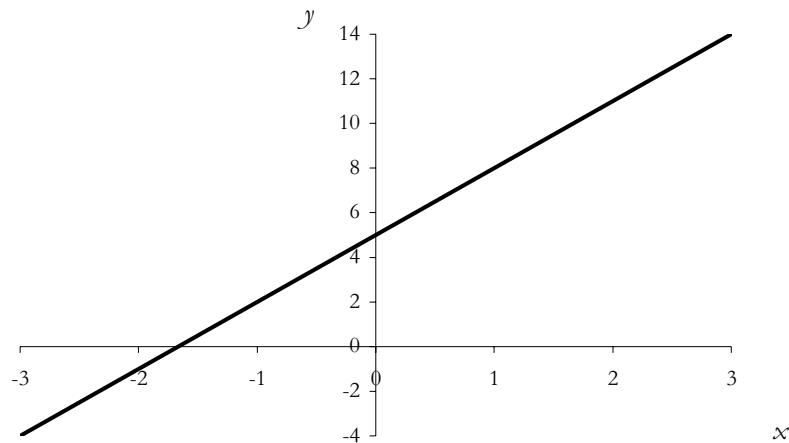
We begin with a brief look at various notations and their uses. Letters (x, y , etc.) are often used in mathematics or statistics to represent numbers. These ‘substitute values’ are known as ‘variables’. Variables are ‘wild card’ characters which can take variable values. In different circumstances letters may be standing for variables or constants. For instance, in the expression $3x + 5$ ($x \in \mathbf{R}$), x is a variable. There are a range of possible values which it can take. In this case the ‘ $x \in \mathbf{R}$ ’ tells us that x can be any real number. In most cases the context will determine the values that a variable can take. The expression is shorthand for $(3 \times x) + 5$, so that if $x = -3$ then the value of the expression is $(3 \times (-3)) + 5 = -4$.

We call $3x + 5$ an *expression*. $f(x) = 3x + 5$ defines a *function* f . Strictly we also need to know the *domain*, i.e. the set of possible x values, such as $x \in \mathbf{R}$, before the function is defined. A function has the property that, for any value of the input(s) there is a unique value output. Note the possible plurality of inputs – we often operate with functions of more than one variable.

Yet another variant on the theme is $y = 3x + 5$. Expressed in this way, x is the *independent* variable and y , computed from a given value of x , is known as the *dependent* variable. Strictly, $y = 3x + 5$ is the *graph* of the function $f(x) = 3x + 5$. It is a graph because for every value given to the variable x , a value for the variable y can be computed. Thus it represents a set of (x, y) pairs, such as $(-3, -4)$, all of which have the property that the y coordinate is 3 times the x coordinate plus 5. We can plot all of these (x, y) pairs. Together they form a line (Figure II.A.1).

¹ Keith Parramore is Principal Lecturer in Mathematics at the University of Brighton and Terry Watsham is Principal Lecturer at the University of Brighton.

Figure II.A.1: Graph of a straight line



Of course, the choice of 3 and 5 in the expression/function/graph was entirely arbitrary. We could just as well have chosen to look at $2x - 4$, or $-5x + 0.5$, or We can express that thought by writing $mx + c$, or $f(x) = mx + c$, or $y = mx + c$. The letters m and c indicate constants, but constants that we can change. They are often referred to as *parameters*.

Note that in the graph $y = mx + c$, m is the *gradient* of the line and c is the *intercept*, i.e. the value at which the line cuts the y -axis. The reader should note that the gradient might not appear to be correct if the scales on the axes are different, but that careful counting will resolve the issue. For instance, in Figure II.A.1 the line passes through $(-1\frac{2}{3}, 0)$ and through $(0, 5)$. In doing so it climbs 5 whilst moving across $1\frac{2}{3}$, and 5 divided by $1\frac{2}{3}$ is indeed 3. To find the point $(-1\frac{2}{3}, 0)$ we determined where the line crosses the x -axis. This happens when $y = 0$, so the x value can be found by solving $3x + 5 = 0$. This is an *equation*.

Example II.A.1

Suppose that the return on a stock, R_s , is related to the return on a related index, R_i , by the equation $R_s = -0.0064 + 1.19R_i$. The return to the index is the independent variable. The return to the stock is the dependent variable. Based on this relationship, the expected returns on the stock can now be computed for different market returns. For $R_i = 0.05$ we compute $R_s = -0.0064 + 1.19 \times 0.05 = 0.0531$. Similarly $R_i = 0.10$ gives $R_s = 0.1126$ and $R_i = 0.15$ gives $R_s = 0.1721$. The parameter 1.19 (also called the *coefficient* of R_i) is the *beta* of economic theory. It measures the sensitivity between stock market returns and stock or portfolio returns, as in the capital asset pricing model (see Chapter I.A.4).

Greek letters are often used to label specific parameters. The table below shows common usages in finance:

Lower-case letter	Upper-case letter	Pronunciation	Examples in finance
α	A	Alpha	Regression intercept
β	B	Beta	Regression slope – systematic risk
γ	Γ	Gamma	Sensitivity measurement for options
δ	Δ	Delta	Sensitivity measurement for options
ϵ	E	Epsilon	Random error or disturbance for regressions
θ	Θ	Theta	Sensitivity measurement for options
κ	K	Kappa	Kurtosis
μ	M	Mu	Expected return
ν	N	Nu	Integer constant, e.g. degrees of freedom
π	Π	Pi	Circle constant – symbol for a product
ρ	P	Rho	Correlation coefficient
σ	Σ	Sigma	Standard deviation – symbol for a sum
τ	T	Tau	Time to maturity – maturity date
φ	Φ	Phi	Normal distribution
χ	X	Chi	Statistical distribution for testing ‘fit’

II.A.1.2 The Algebra of Number

We have all studied algebra, but who could give a definition of what ‘algebra’ actually is? An algebra consists of a set of objects together with a set of operations on those objects, an operation being a means of combining or transforming objects to give a new ‘output’ object. (From the language here you may guess that a more formal definition of ‘operation’ would involve the word ‘function’.) There are many algebras, and they each have their rules for how objects and operations interact. The algebra of logic involves combining statements which can be true or false, using operations such as AND and OR. The algebra of matrices will be covered in Chapter II.D. But of course, when we talk of ‘algebra’ we are usually referring to the algebra of number.

We will not be too formal in what follows. We will not give the complete, exhaustive list of the axioms (rules) of the real number system, nor dwell for too long on the different types of number, but an outline should help jog those memories so carefully laid down during the years of formal education.

Types of Number

There is a hierarchy of number type. We start with the natural numbers, $\{1, 2, 3, 4, 5, \dots\}$, labelled **N**.

Next come the integers, $\mathbf{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$. Note that the integers contain the natural numbers.

After the integers come the rationals (fractions). These are labelled **Q**. They each consist of a pair of integers, the numerator and the denominator. The denominator cannot be zero, and there is an extra element of structure which defines $1/2, 2/4, 3/6$, etc. as being identical.

Next we have the reals, denoted by **R**. These consist of rationals plus irrationals – numbers which cannot be expressed as the ratio of two integers. Examples of irrationals include π, e , and numbers such as $0.101001000100001\dots$

Finally we have the complex numbers, **C**. These will not concern us on this course.

Operations

We are used to dealing with addition, subtraction, multiplication and division. In fact subtraction is defined in terms of addition. For instance, $6 - 2$ is the number which needs to be added on to 2 to give 6. Similarly, division is defined in terms of multiplication. For instance, $6/2$ is the number which when multiplied by 2 gives 6. Thus we only need to deal with rules relating to addition and multiplication. Furthermore, multiplication within the natural numbers is defined in terms of repeated addition.

Having defined the operations within **N** we need to extend the definitions to sets further up the hierarchy. Much time is spent in schools in dealing with operations on fractions. We summarise the algorithms below.

$$\frac{a}{b} + \frac{p}{q} = \frac{aq + bp}{bq} \quad \text{e.g. } \frac{2}{3} + \frac{1}{6} = \frac{12 + 3}{18} = \frac{15}{18} \text{ which cuts down to } \frac{5}{6}.$$

$$\frac{a}{b} \times \frac{p}{q} = \frac{ap}{bq} \quad \text{e.g. } \frac{2}{3} \times \frac{1}{6} = \frac{2}{18} \text{ which cuts down to } \frac{1}{9}.$$

Rules

The rule defining how the operations interact with numbers are drilled into us throughout our formative years. Here are some of them:

The <i>commutativity</i> of addition	$a + b = b + a$	e.g. $12+4 = 16$ $4+12=16$
The <i>associativity</i> of addition	$a + (b + c) = (a + b) + c$	e.g. $12+(4+3) = 12+7=19$ $(12+4)+3 = 16+3 = 19$
The existence of <i>zero</i>	$a + 0 = a$	
The existence of <i>negatives</i>	$a + (-a) = 0$	(but not in \mathbf{N})
The <i>commutativity</i> of multiplication	$a \times b = b \times a$	e.g. $12 \times 4 = 48$ $4 \times 12 = 48$
The <i>associativity</i> of multiplication	$a \times (b \times c) = (a \times b) \times c$	e.g. $12 \times (4 \times 3) = 12 \times 12 = 144$ $(12 \times 4) \times 3 = 48 \times 3 = 144$
The existence of 1	$a \times 1 = a$	
The existence of <i>inverses</i>	$a \times (a^{-1}) = 1$	(but not in \mathbf{Z})
The <i>distributivity</i> of multiplication over addition (i.e. opening brackets)	$a \times (b + c) = (a \times b) + (a \times c)$	e.g. $12 \times (4+3) = 12 \times 7 = 84$ $(12 \times 4) + (12 \times 3) = 48 + 36 = 84$
[But note that addition does not distribute over multiplication: $12+(4 \times 3) \neq (12+4) \times (12 \times 3)$.]		

From these rules we can deduce such results as $a - (-b) = a + b$ and $(-a) \times (-b) = a \times b$.

II.A.1.3 The Order of Operations

The rules of algebra define the order in which operations are applied. The importance of this can be seen by considering expressions such as $12 \div 4 \div 3$. Is this to mean $12 \div (4 \div 3) = 9$ or $(12 \div 4) \div 3 = 1$?

The solution, as you see above, is to use brackets. However, when complex expressions are written down it is common practice to follow conventions which limit the number of brackets that are needed. This makes the expressions easier to read – but you do need to know the conventions! Thus it is common practice to write down a linear expression in the form $2x + 3$ (understanding that $2x$ is itself shorthand for $2 \times x$). It is accepted convention that this is shorthand for $(2x) + 3$, rather than $2(x + 3)$. In other words, the multiplication is to be done before the addition. So if the expression is to be evaluated when $x = 5$, then the answer is $10 + 3 = 13$, and not $2 \times 8 = 16$. If the latter order of operations is required then brackets must be used, i.e. $2(x + 3)$.

These conventions are summarised in the acronym BIDMAS, which gives the order in which to apply operations, unless brackets indicate otherwise – *B*rackets, *I*ndices (powers), *D*ivisions, *M*ultiplications, *A*dditions, *S*ubtractions.

Thus to evaluate $(2x + 3)^2 + 40 / (5 - x / 2) + 5$ when $x = 6$ proceed thus:

<p>B You need to evaluate the brackets first.</p> <p>I Can't do the squaring yet</p> <p>D $(2x + 3)^2 + 40 / (5 - 3) + 5$</p> <p>M $(12 + 3)^2 + 40 / (5 - 3) + 5$</p> <p>A $15^2 + 40 / (5 - 3) + 5$</p> <p>S $15^2 + 40 / 2 + 5$</p>	<p>iterating ...</p> <p>B now done</p> <p>I $225 + 40 / 2 + 5$</p> <p>D $225 + 20 + 5$</p> <p>M none to do</p> <p>A 250</p>
--	---

II.A.2 Sequences and Series

II.A.2.1 Sequences

A sequence is an ordered countable set of terms. By 'countable' we mean that we can allocate integers to them, so that we have a first term, a second term, etc. In other words there is a list. The fact that they are 'ordered' means that there is a specific defined list. The list may be infinite. A sequence is also called a progression, and there are two progressions of importance in finance:

The Arithmetic Progression

This has a first term which is usually denoted by the letter a . Subsequent terms are derived by repeatedly adding on a fixed number d , known as the *common difference*. Letting u_1 represent the first term, u_2 the second term, etc., we have:

$$\begin{array}{ccccccc}
 u_1 & u_2 & u_3 & & u_n & & \\
 a & a + d & a + 2d & \dots & a + (n - 1)d & \dots &
 \end{array}$$

The Geometric Progression

This has a first term also usually denoted by a . Subsequent terms are derived by repeatedly multiplying by a fixed number r , known as the *common ratio*:

$$\begin{array}{ccccccc}
 u_1 & u_2 & u_3 & & u_n & & \\
 a & ar & ar^2 & \dots & ar^{n-1} & \dots &
 \end{array}$$

Thus the 20th term in the arithmetic progression (AP) 1000, 1025, 1050, ... is $1000 + (19 \times 25) = 1475$. The 20th term in the geometric progression (GP) 1000, 1050, 1102.50, ... is $1000 \times 1.05^{19} = 2526.95$.

II.A.2.2 Series

A series is the accumulated sum of a sequence. The commonly-used notation is $S_n =$ the sum of the first n terms (‘sum to n terms’) of the progression, and, for an infinite progression, $S_\infty =$ the limit of the sum to n terms as n increases (if that limit exists).

The Sum to n Terms of an AP

This is generally attributed to Gauss, who was a mathematical prodigy. It is said that, as an infant, he was able instantly to give his teacher the correct answer to problems such as $1 + 2 + 3 + \dots + 100$. This is, of course an AP, and its sum can easily be obtained by writing down two copies of it, the second in reverse:

	1	2	3	...	98	99	100
	100	99	98	...	3	2	1
sum	101	101	101	...	101	101	101

Note that in the representation above there are 100 terms, each summing to 101. Since we have written down the sum twice, it is exactly double the sum of one series. Thus

$$1 + 2 + 3 + \dots + 100 = \frac{100 \times 101}{2} = 50,500.$$

This generalises easily to give a formula for the sum to n terms of an AP (note that the sum to infinity of an AP is infinite, so we just take the sum up to a finite number of terms as follows). Let $S_n = a + (a + d) + (a + 2d) + \dots + (a + (n - 1)d)$. Then

$$S_n = \frac{n}{2}(2a + (n - 1)d). \tag{II.A.1}$$

For instance, $1000 + 1025 + 1050 + \dots + 1475 = 10(2000 + 19 \times 25) = 24,750$.

The Sum to n Terms of a GP

To find $S_n = a + ar + ar^2 + \dots + ar^{n-1}$, multiply through by r and write it below:

S_n	a	ar	ar^2	...	ar^{n-3}	ar^{n-2}	ar^{n-1}	
rS_n		ar	ar^2	...	ar^{n-3}	ar^{n-2}	ar^{n-1}	ar^n
difference	a			...				$- ar^n$

Thus $S_n - rS_n = a - ar^n$, giving:

$$S_n = \frac{a(1 - r^n)}{1 - r} \text{ if } r < 1 \quad \text{and} \quad S_n = \frac{a(r^n - 1)}{r - 1} \text{ if } r > 1 \tag{II.A.2}$$

And, of course, $S_n = na$ if $r = 1$.

For instance,

$$1000 + 1050 + 1102.50 + \dots + 2526.95 = \frac{1000 \times (1.05^{20} - 1)}{1.05 - 1} \approx 33,065.95.$$

For a GP, if $-1 < r < 1$ (i.e. if $|r| < 1$) then S_n *does* tend to a limit as n tends to infinity. This is because the r^n term tends to zero, so that:

$$S_\infty = \frac{a}{1-r}. \tag{II.A.3}$$

II.A.3 Exponents and Logarithms

II.A.3.1 Exponents

The fundamental rule of exponents (powers) is abstracted from simple observations such as $2^3 \times 2^2 = 2^5$, since $2^3 = 2 \times 2 \times 2 = 8$, $2^2 = 2 \times 2 = 4$ and $2^5 = 2 \times 2 \times 2 \times 2 \times 2 = 32$. The generalization of this is known as the first rule of exponents: For $a > 0$,

- $a^x \times a^y = a^{x+y}$ (This may *seem* obvious, but it applies to all numbers x and y , not just to natural numbers.)

There are two further rules of exponents. They are not really *needed* since they follow from rule 1, but they are useful to have:

- $a^x \div a^y = a^{x-y}$
- $(a^x)^y = a^{xy}$ (II.A.4)

The consequences of these rules are immediate:

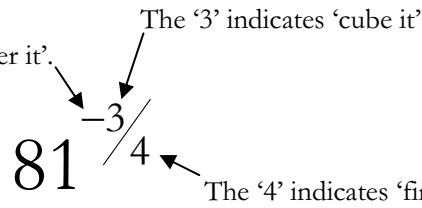
$$a^0 = 1 \text{ for all } a. \quad \text{For instance, } 2^0 = 1 \text{ because, as an example, } 2 \times 2^0 = 2^1 \times 2^0 = 2^{1+0} = 2.$$

$$a^{-1} = \frac{1}{a} \quad \text{For instance, } 2^{-1} = \frac{1}{2} \text{ because, as an example, } 2 \times 2^{-1} = 2^{1+(-1)} = 2^0 = 1.$$

$$a^{1/2} = \sqrt{a} \quad \text{For instance, } 2^{1/2} = \sqrt{2} \text{ because, as an example, } 2^{1/2} \times 2^{1/2} = 2^{1/2+1/2} = 2^1 = 2.$$

So an exponent may have three parts. For instance:

The negative sign indicates the multiplicative inverse, i.e. ‘one over it’.



These three operations can be done in any order. In this case it is easiest to take the fourth root

first, since $81 = 2^4$. So $81^{-3/4} = \frac{1}{2^3} = \frac{1}{8}$.

II.A.3.2 Logarithms

A logarithm is the inverse of an exponent. For instance:

$$\log_{10}(100) = 2 \text{ because } 10^2 = 100;$$

$$\log_{10}(1000) = 3 \text{ because } 10^3 = 1000;$$

$$\log_{10}(1,000,000) = 6 \text{ because } 10^6 = 1,000,000;$$

$$\log_2(81) = 4 \text{ because } 2^4 = 81;$$

$$\log_5(0.04) = -2 \text{ because } 5^{-2} = \frac{1}{25} = 0.04.$$

This is all encapsulated in the following defining relationship:

$$\log_a(x) = y \Leftrightarrow x = a^y. \tag{II.A.5}$$

It says that $\log_a(x)$ is that power to which a must be raised to give the answer x .

Corresponding to the three rules of exponents are three rules of logarithms. They are:

1. $\log_a(xy) = \log_a(x) + \log_a(y)$
2. $\log_a(x/y) = \log_a(x) - \log_a(y)$
3. $\log_a(x^y) = y \log_a(x)$ (II.A.6)

We demonstrate the correspondence in the case of rule 1, showing in the process how to use equation (II.A.5):

$$\text{Let } p = \log_a(x) \text{ and } q = \log_a(y).$$

$$\text{Then } x = a^p \text{ and } y = a^q \text{ (by equation (II.A.5)).}$$

So $xy = a^p \times a^q = a^{p+q}$ (by the first rule of exponents).

So $p + q = \log_a(xy)$ (by equation (II.A.5)).

That is, $\log_a(x) + \log_a(y) = \log_a(xy)$.

Suppose that we needed to find $\log_2(6)$. Since $2^2 = 4$ and $2^3 = 8$ it must be between 2 and 3. But tables of logarithms only ever use base 10 and base e (where e is a very important, irrational, constant with approximate value 2.71828182845905 – see below). However, we can again use equation (II.A.5). Let $x = \log_2(6)$. Then $2^x = 6$. Now taking logs (base 10) of both sides gives:

$$\log_{10}(2^x) = \log_{10}(6), \text{ i.e. } x \log_{10}(2) = \log_{10}(6), \text{ so } x = \frac{\log_{10}(6)}{\log_{10}(2)} \approx 2.585.$$

Note that in Excel you can find $\log_2(6)$ directly from the formula ‘=log(6,2)’.

We will see practical applications of this in Section II.A.6, when we look at continuous compounding. Pre-empting that work to some extent, consider a sum of money growing at 4% per annum, continuously compounded. How long will it take for the principal to double? We will see in Section II.A.6 that over time t (years) the principal will grow from P to $Pe^{0.04t}$, so we require t such that $e^{0.04t} = 2$. Using equation (II.A.5), this is the same as $0.04t = \log_e(2)$. $\log_e(2)$ is often written as $\ln(2)$ (the natural log of 2). It has value 0.693147 (from ‘=ln(2)’ in Excel). So the time required to double your money is $0.693147/0.04 = 17.3287$ years.

II.A.3.3 The Exponential Function and Natural Logarithms

In the section above you encountered the expression 2^x . This is a *function* of x – the exponential function to the base 2. All such exponential functions can be transformed into other bases by using the laws of exponents. For example, $2^x = (3^{\log_3(2)})^x = 3^{x \log_3(2)}$. However, there is one *natural* base which acts as the ‘gold standard’ for exponents – ‘e’. This is an irrational number.

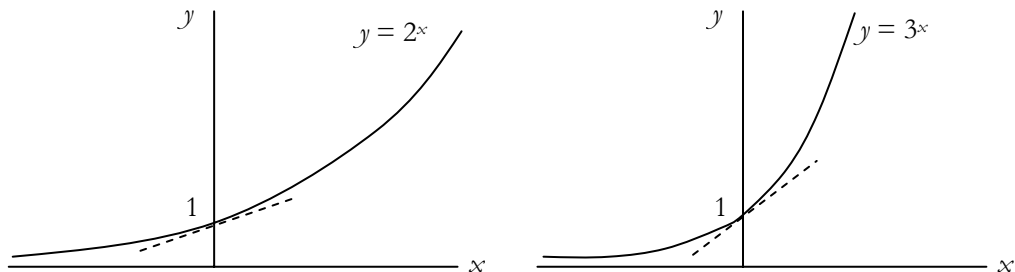
What is e?

To 500 significant figures its value is:

e = 2.71828182845904523536028747135266249775724709369995957496696762772
 407663035354759457138217852516642742746639193200305992181741359
 662904357290033429526059563073813232862794349076323382988075319
 525101901157383418793070215408914993488416750924476146066808226
 480016847741185374234544243710753907774499206955170276183860626
 133138458300075204493382656029760673711320070932870912744374704
 723069697720931014169283681902551510865746377211125238978442505
 69536967707854499699679468644549059879316368892300987931

The obvious question is ‘What is ‘natural’ about e ?’. Answering this is a precursor to work later in this chapter (Section II.A.5.3) and later still in Chapter II.C. Figure II.A.2 shows sketch graphs of $y = 2^x$ and $y = 3^x$.

Figure II.A.2: Graphs of 2^x and 3^x



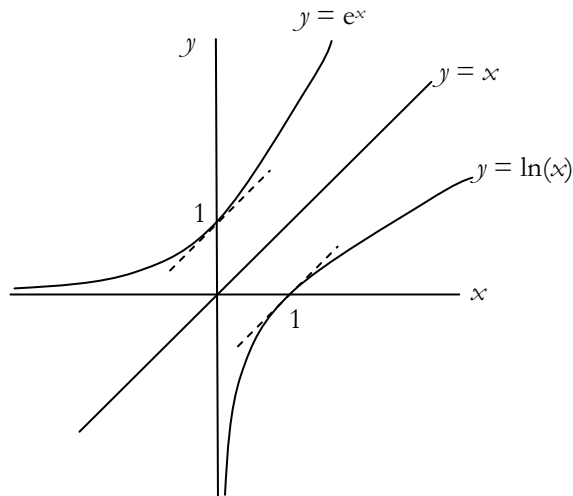
The second curve is below the first for negative values of x and above it for positive values of x . Its slope at $x = 0$ (indicated by the dotted line) is greater than 1. The slope of $y = 2^x$ at $x = 0$ is less than 1. Thus there must be some base, between 2 and 3, for which the slope of the graph at $x = 0$ is exactly 1. This base is e .

This defining property has enormous significance in mathematics, and in financial mathematics. One of the consequences is that e^x is the only function which differentiates to itself, and a consequence of that is that the integral of $1/x$ is $\log_e(x)$. Chapter II.C lists the common rules for differentiation and integration, including these rules of the exponential and \log_e functions.

What is the natural logarithm?

The natural logarithm is the function $\log_e(x)$. That is, it is the inverse of e^x . Being the inverse, its graph must be the reflection of $y = e^x$ in the 45 degree line $y = x$, as shown in Figure II.A.3.

Figure II.A.3: Graphs of $\exp(x)$ and $\ln(x)$



Thus, for instance, $\log_e(e^2) = 2 \log_e(e) = 2$ and $e^{\log_e(2)} = 2$.

Note that we normally employ alternative, special notations for the two functions:

$$\exp(x) = e^x \text{ and } \ln(x) = \log_e(x),$$

so $\exp(\ln(2)) = 2$. Note also, from Figure II.A.3, that $\exp(0) = 1$ and that $\ln(1) = 0$.

Expansions

Both functions can be expressed as infinite sums (see Section II.A.2):

$$\exp(x) = 1 + \frac{x}{1} + \frac{x^2}{2} + \frac{x^3}{6} + \dots \quad (\text{II.A.7})$$

and

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots \text{ provided } -1 < x \leq 1. \quad (\text{II.A.8})$$

These expressions become very useful for approximations. For instance, from the above we know that

$$\text{for small values of } x, \ln(1+x) \approx x. \quad (\text{II.A.9})$$

This result is of particular importance in considering the return to an investment over a short time period (see Section II.A.6). The discretely compounded return is

$$\frac{Y(t+\delta t)}{Y(t)} - 1,$$

where $Y(t)$ is the value of the investment at time t . The continuously compounded return is given

$$\text{by } \ln\left(\frac{Y(t + \delta t)}{Y(t)}\right).$$

But since $\ln(1+x) \approx x$ the continuously compounded return is approximately equal to the discretely compounded return, provided the discretely compounded return is small. Thus, for small time periods, the discretely compounded rate of return and the continuously compounded rate of return are approximately equal.

Similarly, we can use $\ln(1+x) \approx x$ to show that, if P_t is a price measured at time t , then

$$(P_{t+1} - P_t)/P_t \approx \ln(P_{t+1}/P_t) \tag{II.A.10}$$

and the right hand side is $\ln(P_{t+1}) - \ln(P_t)$. Hence it is common to approximate discretely compounded returns over short periods by the difference in the log prices.

II.A.4 Equations and Inequalities

II.A.4.1 Linear Equations in One Unknown

In financial market theory the return on a security is often regressed on the market return, i.e. the return on securities is viewed as being *linearly dependent* (see Chapter II.D) on the market return. The coefficient of the market return represents the so-called beta factor (or β). The intersection with the axis is the market-independent return α . This is the capital assets pricing model

$$R = \alpha + \beta R_M,$$

where R is the stock return (dependent variable), R_M the market return (independent variable), α the market-independent return (constant) and β the sensitivity measurement (coefficient).

Suppose that historic data give $\alpha = 0.002965$ and $\beta = 1.0849$, so that $R = 0.002965 + 1.0849R_M$. Then if the market return is 10%, we can expect a return on the portfolio of $R = 0.002965 + 1.0849 \times 0.10 = 0.111455$.

The question now is, what is the market return if the portfolio return is, say, 10%? To answer this, $0.10 = 0.002965 + 1.0849R_M$ must be solved in terms of R_M .

This is an example of a linear equation in one unknown. It is solved by applying a sequence of operations to both sides of the equation, thus preserving the equality, until the unknown is isolated. In this case:

Step 1 Subtract 0.002965 from both sides $0.10 - 0.002965 = 1.0849R_M$

Step 2 Divide both sides by 1.0849 $(0.10 - 0.002965)/1.0849 = R_M$

So for a portfolio return $R = 10\%$, the market return is $R_M = (0.10 - 0.002965)/1.0849 = 0.089441$, i.e. 8.9441%.

II.A.4.2 Inequalities

Linear inequalities are handled in exactly the same way as linear equalities, *except* in that when both sides are multiplied or divided by a negative number, the direction of the inequality must be changed. Thus, to solve $3 - 2x < 9$:

Step 1 Subtract 3 from both sides $-2x < 6$

Step 2 Divide both sides by -2
and reverse the direction of the inequality $x > -3$

Step 3 Check the solution:

e.g. $x = -2.5$ is in the solution set because $3 - 2 \times (-2.5) = 8$, and $8 < 9$.

The need to reverse the direction of the inequality can always be avoided. It is worth seeing what happens when it is avoided to understand the need for the extra rule:

Step 1 Add $2x$ to both sides $3 < 2x + 9$

Step 2 Subtract 6 from both sides $-6 < 2x$

Step 3 Divide both sides by 2 $-3 < x$

II.A.4.3 Systems of Linear Equations in More Than One Unknown

We often have to solve an *equation system* of two variables (e.g. x and y); that is, we have to determine numerical values for both variables so that both of the original equations are satisfied. Here we restrict ourselves to two equations in two unknowns. In Chapter II.D you will find examples of systems with more than two variables and more than two equations.

Suppose:

$$a_1x + b_1y + c_1 = 0 \quad (1\text{st equation})$$

$$a_2x + b_2y + c_2 = 0 \quad (2\text{nd equation})$$

where x and y are variables, a_1 , a_2 , b_1 and b_2 are coefficients, and c_1 and c_2 are constants. Values must now be found for x and y to satisfy both equations at the same time.

The Elimination Method

Step 1 Both equations are solved in terms of one of the variables (e.g. y)

$$a_1x + b_1y + c_1 = 0 \Rightarrow y = (-a_1x - c_1)/b_1$$

$$a_2x + b_2y + c_2 = 0 \Rightarrow y = (-a_2x - c_2)/b_2$$

Step 2 Equate the two expressions for y

$$(-a_1x - c_1)/b_1 = (-a_2x - c_2)/b_2$$

Step 3 Solve in terms of x

$$x(a_2/b_2 - a_1/b_1) = c_1/b_1 - c_2/b_2 \Rightarrow x = (c_1b_2 - c_2b_1)/(a_2b_1 - a_1b_2)$$

Step 4 Determine y

The value of y can be determined by substituting the value calculated for x (under step 3) into one of the original equations.

Example II.A.2

Solve: $4y - 2x - 3 = 0$
 $2y + 3x - 6 = 0$

Note that $a_1 = -2$, $b_1 = 4$, $c_1 = -3$, $a_2 = 3$, $b_2 = 2$ and $c_2 = -6$.

Solution:

$$x = \frac{-6 - (-24)}{12 - (-4)} = 1.125 \text{ and } y = 1.3125$$

The Substitution Method

In the substitution method, one of the two equations is solved in terms of one of the two variables, and the result is substituted in the other equation. The approach is illustrated using the above example, rewritten as:

$$\begin{aligned} -2x + 4y - 3 &= 0 && \text{(1st equation)} \\ 3x + 2y - 6 &= 0 && \text{(2nd equation)} \end{aligned}$$

Step 1 Solve the first equation in terms of x :

$$\begin{aligned} 2x &= 4y - 3 \\ x &= (4y - 3)/2 = 2y - 1.5 \end{aligned}$$

Step 2 Substitute in the second equation.

$$\begin{aligned} 3x + 2y - 6 &= 0 \Rightarrow 3(2y - 1.5) + 2y - 6 = 0 \\ 6y - 4.5 + 2y - 6 &= 0 \\ 8y - 10.5 &= 0 \\ 8y &= 10.5 \\ y &= 1.3125 \end{aligned}$$

Step 3 Determine the value of x :

$$x = 2y - 1.5 = 2 \times 1.3125 - 1.5 = 1.125$$

II.A.4.4 Quadratic Equations

A quadratic equation has the structure $ax^2 + bx + c = 0$, where x is the *variable*, a and b are *coefficients* and c is a *constant*. The question is, which values of x satisfy the equation? A quadratic equation has a maximum of two solutions, which can be found using the following formula:

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \text{ and } x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \quad (\text{II.A.11})$$

Note that the solution depends on the quantity $b^2 - 4ac$, which is known as the *discriminant*:

- If the discriminant is negative then there are no real solutions.
- If the discriminant is zero there is only one solution, $x = -b/2a$.
- If the discriminant is positive there are two different real solutions.

Example II.A.3

Solve $2x^2 + 3x - 4 = 0$.

$$x = \frac{-3 \pm \sqrt{3^2 - 4 \times 2 \times (-4)}}{2 \times 2} = \frac{-3 \pm \sqrt{9 + 32}}{4} = \frac{-3 \pm \sqrt{41}}{4} = \frac{-3 \pm 6.4031}{4}$$

$$\text{So } x_1 = \frac{-3 + 6.4031}{4} = 0.8508 \text{ and } x_2 = \frac{-3 - 6.4031}{4} = -2.3508.$$

Example II.A.4

Solve $2x^2 + 3x + 4 = 0$.

$$x = \frac{-3 \pm \sqrt{3^2 - 4 \times 2 \times 4}}{2 \times 2} = \frac{-3 \pm \sqrt{9 - 32}}{4} = \frac{-3 \pm \sqrt{-23}}{4}$$

In this case the discriminant is negative, so there are no real solutions.

Example II.A.5

Solve $2x^2 + 4x + 2 = 0$.

$$x = \frac{-4 \pm \sqrt{4^2 - 4 \times 2 \times 2}}{2 \times 2} = \frac{-4 \pm \sqrt{0}}{4}$$

$$\text{So } x_1 = x_2 = -0.75.$$

II.A.5 Functions and Graphs

II.A.5.1 Functions

A *function* is a rule that assigns to every value of x exactly one value of y . A function is defined by its *domain* (the set of elements on which it operates) together with its action (what it does). The latter is usually specified by a rule, but for finite and small domains it may be specified by a list. The domain is often not explicitly stated since it can often be inferred from the context.

Example II.A.6

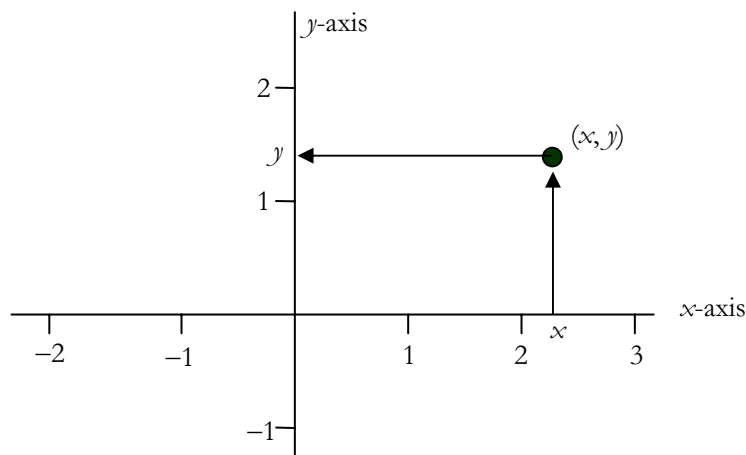
1. $f(x) = 3x + 5$ ($x \in \mathbf{R}$) This is a linear function (see Section II.A.1.1). The bracket specifies the domain. It reads ‘ x is a member of \mathbf{R} ’; i.e. it says that x can be any real number.
2. $f(x) = 2x^2 + 3x - 4$ ($x \in \mathbf{R}$) This is a quadratic function (see Section II.A.4.4).
3.

x	1	2	3	4	Here the function p is defined by a list.
$p(x)$	0.1	0.3	0.5	0.1	(This is an illustration of a random variable – see Chapter II.E.)

II.A.5.2 Graphs

The *graph* of a function f is obtained by plotting points defined by all values of x and their associated images $f(x)$, as shown in Figure II.A.4. The x values are measured along the horizontal axis or x -axis (abscissa). The associated function values are measured along the vertical axis or y -axis (ordinate). Thus $y = f(x)$ is known as the graph of f .

Figure II.A.4



II.A.5.3 The Graphs of Some Functions

Figure II.A.5: [Graph of \$y = 3x + 5\$](#)

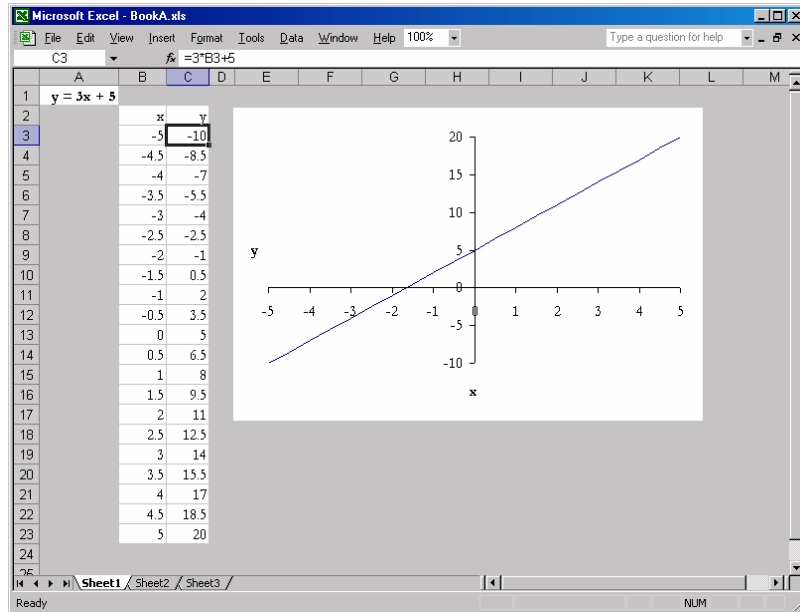


Figure II.A.5 shows the graph of $y = 3x + 5$, which is a straight line with slope 3 and intercept 5. It is exactly the same as Figure II.A.1. The graph in Figure II.A.6 shows the first two quadratic functions from Section II.A.4.4. The graph $y = 2x^2 + 3x - 4$ is the solid line and $y = 2x^2 + 3x + 4$ is the dashed line. It shows why there were two solutions to the equation $2x^2 + 3x - 4 = 0$ (at $x = 0.8508$ and at $x = -2.3508$) and no solutions to the equation $2x^2 + 3x + 4 = 0$. Figures II.A.7 shows the graph of $y = e^x$ and Figure II.A.8 the graph of $y = \log_e(x)$ – see also Figure II.A.3.

Figure II.A.6: [Graph of \$y = 2x^2 + 3x - 4\$ and \$y = 2x^2 + 3x + 4\$](#)

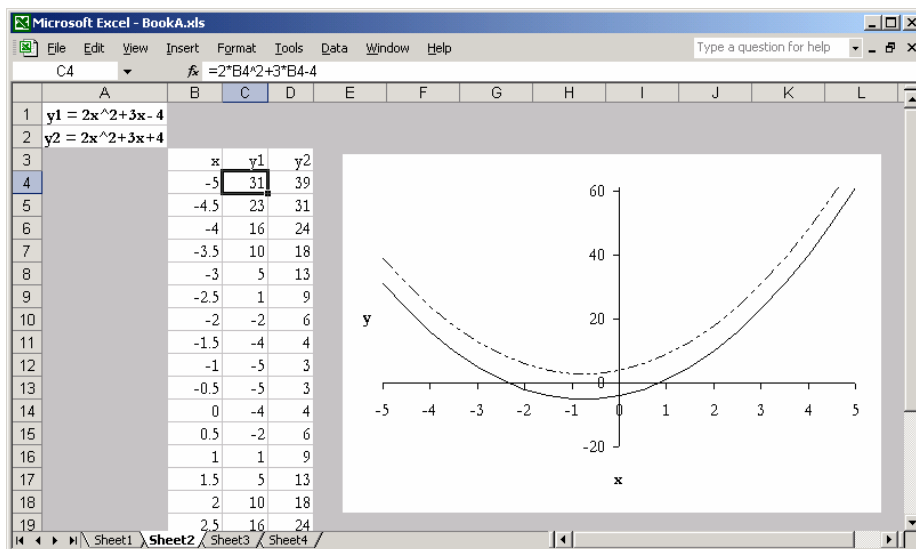


Figure II.A.7: Graph of $y = e^x$

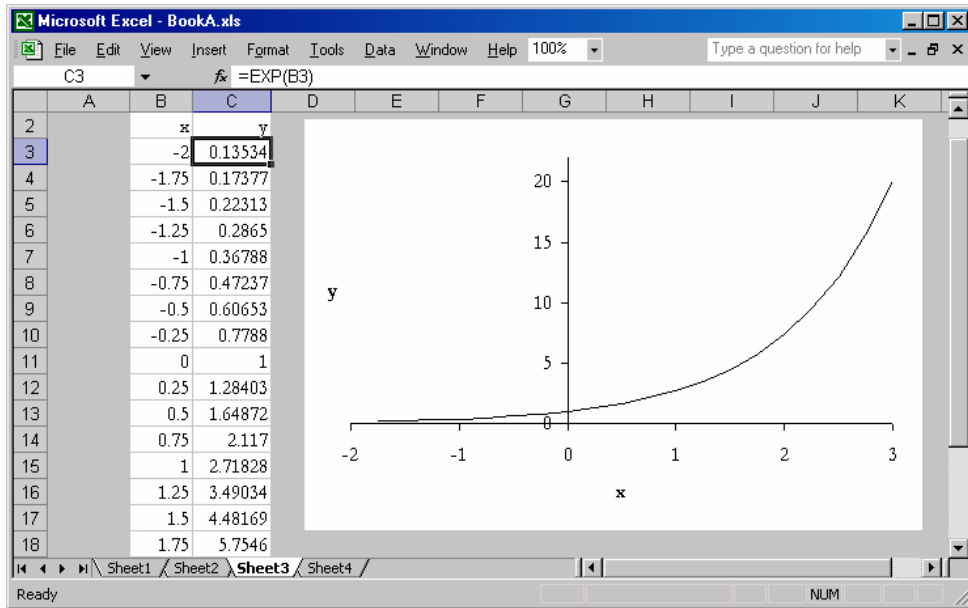
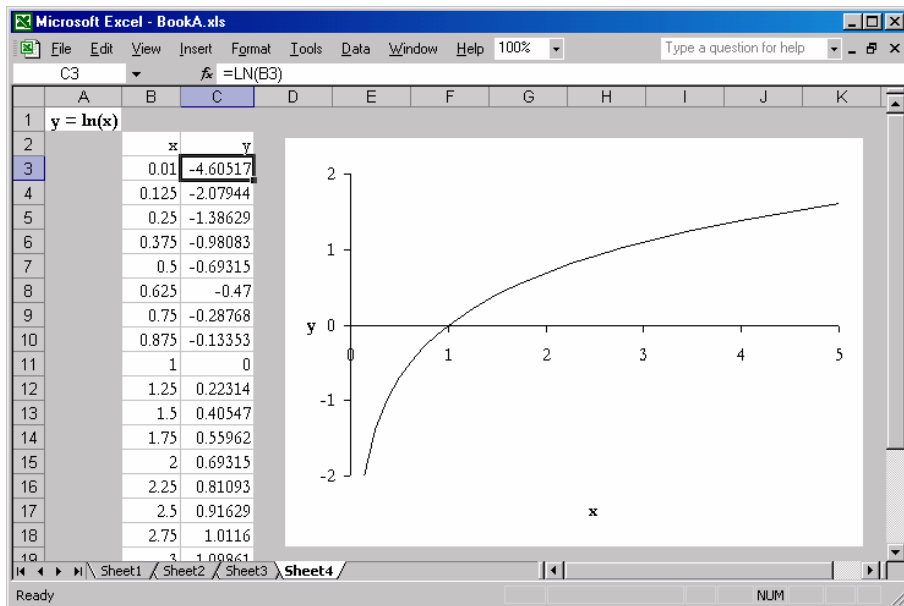


Figure II.A.8: Graph of $y = \log_e(x)$



II.A.6 Case Study – Continuous Compounding

II.A.6.1 Repeated Compounding

Consider the following problem: An investor has invested CHF 100,000 in a bank savings account. The interest rate on the account concerned is 4% per annum. To what sum does his investment grow in 3 years, given that the interest rate does not vary in the next three years and that the interest is credited to the account at the end of each year? The answer is $100,000 \times 1.04^3 = \text{CHF } 112,486.40$. The 1.04 is known as the *compounding factor*. Multiplying by it adds on 4% interest.

Now suppose that the interest is credited (i.e. *compounded*) twice a year. Then interest of 2% must be credited six times, so the final amount would be $100,000 \times 1.02^6 = \text{CHF } 112,616.24$.

In this section we consider what happens as the compounding interval continues to be reduced. To see the effect of this we look at the result of investing CHF 100,000 for one year at different interest rates, and under different compounding regimes. The results are shown in Figure II.A.9.

Figure II.A.9: Final values from investing CHF 100,000 for one year

The screenshot shows a Microsoft Excel spreadsheet titled 'Compounding.xls'. The formula bar displays the formula: $=100000*(1+G\$1/(\$B8*100))^{\$B8}$. The spreadsheet contains the following data:

	A	B	C	D	E	F	G
1		Interest rate (%)	2	5	10	20	100
2			102000.00	105000.00	110000.00	120000.00	200000.00
3	Number of	2	102010.00	105062.50	110250.00	121000.00	225000.00
4	compounding	4	102015.05	105094.53	110381.29	121550.63	244140.63
5	intervals per	12	102018.44	105116.19	110471.31	121939.11	261303.53
6	annum	365	102020.08	105126.75	110515.58	122133.59	271456.75
7		8760	102020.13	105127.09	110517.03	122140.00	271812.67
8		525600	102020.13	105127.11	110517.09	122140.27	271827.92
9							

The ‘number of compounding periods per annum’ represents annual, biannual, quarterly, daily, hourly and minute-by-minute compounding. For each rate of interest, i.e. for each column, you can see that as the number of compounding periods increase (i.e. as the compounding intervals decrease) the final value increases, but that in each case it increases towards a limit.

To see what those limits are, consider Figure II.A.10.

Figure II.A.10: [Comparisons with the limiting values](#)

	A	B	C	D	E	F	G
1		Interest rate (%)	2	5	10	20	100
2		1	102000.00	105000.00	110000.00	120000.00	200000.00
3	Number of	2	102010.00	105062.50	110250.00	121000.00	225000.00
4	compounding	4	102015.05	105094.53	110381.29	121550.63	244140.63
5	intervals per	12	102018.44	105116.19	110471.31	121939.11	261303.53
6	annum	365	102020.08	105126.75	110515.58	122133.59	271456.75
7		8760	102020.13	105127.09	110517.03	122140.00	271812.67
8		525600	102020.13	105127.11	110517.09	122140.27	271827.92
9							
10			102020.13	105127.11	110517.09	122140.28	271828.18
11							

The entry in cell G10 is 100,000 times the mathematical constant e (see Section II.A.3.3). You will see more of the derivation of this number in Chapter II.C. It is a physical constant which is of extreme importance in mathematics, precisely because it underpins the results of many limiting processes. The entries in cells C10 to F10 are $100,000e^{0.02}$, $100,000e^{0.05}$, $100,000e^{0.10}$ and $100,000e^{0.20}$, respectively.

Finally, to see what happens if we invest for more than one year consider Figure II.A.11. This shows the final value of investing CHF 100,000 for three years at the different interest rates and with minute-by-minute compounding. These are compared to $e^{0.06}$, $e^{0.15}$, $e^{0.30}$, $e^{0.60}$ and $e^{3.00}$, respectively.

Figure II.A.11: [Final values from investing CHF 100,000 for three years](#)

	A	B	C	D	E	F
1	Interest rate (%)	2	5	10	20	100
2						
3	compounding					
4	min-by-min	106183.65	116183.42	134985.88	182211.86	2008547.96
5	continuously	106183.65	116183.42	134985.88	182211.88	2008553.69
6						

The limiting compounding regime is known as *continuous compounding*. The corresponding mathematics is continuous mathematics, the techniques of which are the differential and the integral calculus (see Chapter II.C). This is the compounding regime that is applied in derivatives markets.

The formula for continuous compounding

To sum up, you can see that a sum of money, P , invested for time t (years) at an annual interest rate r (expressed as a decimal) will grow to

$$Pe^{rt}. \tag{II.A.12}$$

II.A.6.2 Discrete versus Continuous Compounding

We have already seen (Section II.A.3.2) that it takes 17.33 years for an investment to double in value when it is growing at 4% per annum continuously compounded. That was given by solving

$$Pe^{0.04t} = 2P,$$

the left-hand side being an application of expression (II.A.12).

If the interest is compounded annually rather than continuously, then the computation is more difficult. The formula for discrete compound growth is:

$$P\left(1 + \frac{r}{n}\right)^m \tag{II.A.13}$$

where r is the interest rate in decimals, n is the number of compounding intervals per annum, and m is the total number of compounding periods. But note that m must be a whole number. If we use logs to solve $P(1.04)^m = 2P$ (noting that $n = 1$ for annual compounding), then we end up with $m = \log(2)/\log(1.04) \approx 17.673$. This is not a whole number, and is therefore not the answer, but at least tells us that 17 years is not long enough and that 18 years is too long.

To find out how far into the 18th year we have to wait before our money is doubled we have to do a separate calculation. We must compute how much is in the account at the end of the seventeenth year, and then see what proportion of the year is required for interest to accumulate to the balance. Letting that proportion be x , the equation we have to solve is:

$$1.04^{17} P \times 0.04 \times x = 2P - 1.04^{17} P.$$

This solves to

$$x = \frac{2 - 1.04^{17}}{1.04^{17} \times 0.04} \approx 0.6687.$$

So it takes 17.6687 years to double your money under annual compounding at 4% per annum.

II.A.7 Summary

Much of what has been covered in this chapter may be second nature, given the mathematical training to which we are all subjected. However, it will be useful to have revisited the material, and to have thought more about it, before plunging into the rest of mathematical finance. If the foundations are solid the building will stand!

II.B Descriptive Statistics

Keith Parramore and Terry Watsham¹

II.B.1 Introduction

In this chapter we will use descriptive statistics to describe, or summarise, the historical characteristics of financial data. In the chapter on probability (Chapter II.E), you will see comparable measures, known as ‘parameters’ rather than ‘descriptive statistics’. Parameters describe the expected characteristics of a random variable over a future time period. Statistics and probability go hand in hand: we try to make sense of our observations so that we may understand and perhaps control the mechanism that creates them. The first step in this is ‘statistical inference’, where we *infer* properties of the mechanism from the available evidence. This is called building a model. We then *deduce* consequences from our model and check to see if they tally with our observations. This process is called *validating* the model. If the model is a good one then it will give us insight and control.

In the context of finance, and in many other contexts, our observations are *stochastic* – there is an element of *chance* (often called *error*) involved. Thus a repeat observation of the same process with all inputs held constant is not guaranteed to produce the same result. Under these circumstances our observations are called *data*, and are usually represented by lower-case Latin letters, such as x_1, x_2, x_3 . The models we produce are ‘probability models’, that is, they are based on the idea of a random variable. This is covered in Chapter II.E. Note that capital letters are used for random variables. Thus we might identify the random variable X with *realisations* x_1, x_2, x_3, \dots that are (for our purposes) real numbers. These ‘realisations’ can represent actual data, that is, values that X has actually taken in a sample, or hypothetical values of X .

Similar random variables differ by virtue of having different defining *parameters*. So, when building a model we have to identify the fundamental characteristics of the appropriate random variable *and* obtain estimates for the parameters. We do this by collecting data and computing *statistics*. These statistics provide estimates of the parameters. We use Latin letters for statistics, for example \bar{x} and s (see Sections II.B.4 and II.B.5 for definitions). Greek letters are used for the corresponding parameters, for example μ and σ (see Chapter II.E for definitions).

In this section the data that we describe will be the historical time series of returns of particular equity and bond indices. We will want to measure an average of the individual returns over a

¹ University of Brighton, UK.

given period. We will also need to measure how dispersed the actual returns are around that average. We may need to know whether the returns are symmetrically distributed around the average value, or whether the distribution is more or less peaked than expected. In addition, we will be interested in how one set of data behaves in relation to other sets of data.

The averages are often referred to as *measures of location* or *measures of central tendency*. The statistics that measure the spread of the data are known as *measures of dispersion*. Statistics that describe the symmetry or otherwise of the data are known as *measures of skewness*, and those that describe the ‘peakedness’ of the data are known as *measures of kurtosis*. The statistics that show how one set of data behaves relative to others are known as *measures of association*.

We will illustrate our examples with the use of quarterly continuously compounded returns from Q1 1994 to Q4 2003 for the MSCI Equity Index and the MSCI Sovereign Bond Index. All data and formulae are given in the accompanying [Excel workbook](#).

II.B.2 Data

Data come in various forms and sometimes the form of the data dictates, or at least influences, the choice of method to be used in the statistical analysis. Therefore, before we begin to investigate the various techniques of statistical analysis, it is appropriate to consider different forms of data.

II.B.2.1 Continuous and Discrete Data

Data may be classified as *continuous* or *discrete*. Discrete data are data that result from a process of counting. For example, financial transactions are discrete in that half or quarter transactions have no meaning. Data relating to asset prices may be discrete due to rules set by individual markets regarding price quotations and minimum price movements. For example, in the UK, government bonds are quoted in thirty-seconds (1/32) of a point or pound. Consequently price data are discrete, changing only by thirty-seconds or multiples thereof. Markets for bonds, equities, futures and options are other examples where minimum price change rules cause the data to be discrete.

Continuous data are data that can take on any value within a continuum, that is, the data are *measured* on a continuous scale and the value of that measurement is limited only by the degree of precision. A typical example is the percentage rate of return on an investment. It may be 10%, or 10.1% or indeed 10.0975%. Data relating to time, distance and speed are other examples of continuous data.

II.B.2.2 Grouped Data

Consider the data in on the quarterly levels and returns of the MSCI World Equity Index from Q1 1994 to Q4 2003. There are 40 observations of the levels data and 39 observations of the returns data. When the data set is large it needs to be summarised in a *frequency table* before the analyst can meaningfully comprehend the characteristics of the data. This gives the data in groups or class intervals.

Creating a Frequency Distribution with Excel

Examine the Excel file. The first worksheet gives the data. In the ‘frequency’ worksheet, the Excel function FREQUENCY creates the frequency distribution (Figure II.B.1). The first step is to create a column of cells that form the set of ‘bins’. Each of these bins represents one class interval. Each bin is defined by the upper limit of the interval. Next ‘select’ the column of cells next to and corresponding to each cell in the array of bins. These cells will receive the frequencies calculated by Excel. Next access the FREQUENCY function. Enter the first and last cell addresses of data observations in the first array. Enter the first and last cell addresses of the array of bins in the second array. Finally, *do not* press the enter button. Instead the data and bins have to be entered as an array formula. To do that, hold down the control and shift keys and simultaneously press enter. The frequency for each class interval should then show in each of the highlighted cells.

When grouping data, attention must be paid to the class intervals:

- Firstly, they should not overlap.
- Secondly, they should be of equal size unless there is a specific need to highlight data within a specific ‘subgroup’, or if data are so limited within a group that they can safely be amalgamated with a previous or subsequent group without loss of meaning.
- Thirdly, the class interval should not be so large as to obscure interesting variation within the group.
- Lastly, the number of class intervals should be a compromise between the detail which the data are expected to convey and the ability of the analyst to comprehend the detail.

The frequency table may well be presented as in Figure II.B.2.

Figure II.B.1

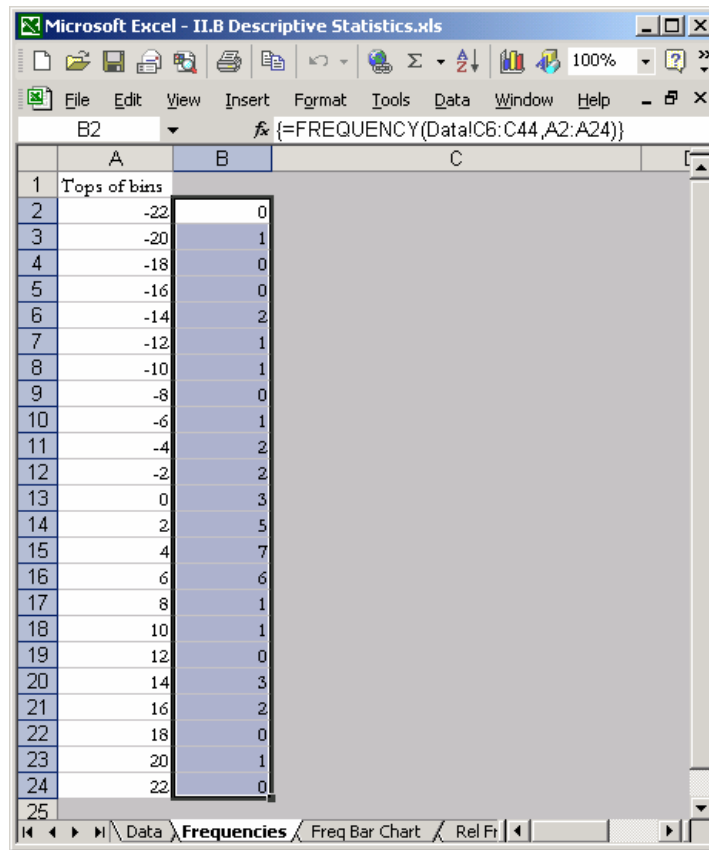


Figure II.B.2

Class Interval	Frequency	Class Interval	Frequency
Up to -22	0	0 and up to +2	5
-22 and up to -20	1	+2 and up to +4	7
-20 and up to -18	0	+4 and up to +6	6
-18 and up to -16	0	+6 and up to +8	1
-16 and up to -14	2	+8 and up to +10	1
-14 and up to -12	1	+10 and up to +12	0
-12 and up to -10	1	+12 and up to +14	3
-10 and up to -8	0	+14 and up to +16	2
-8 and up to -6	1	+16 and up to +18	0
-6 and up to -4	2	+18 and up to +20	1
-4 and up to -2	2	+20 and up to +22	0
-2 and up to 0	3	Over +22	0

In the description of the class intervals, the ‘up to’ statement might be defined as meaning ‘up to but not including’. For example, up to -22 would mean all numbers below -22 . Likewise 0 and up to $+2$ would mean all values including zero and up to but not including $+2$.²

II.B.2.3 Graphical Representation of Data

This section will look at the following ways of presenting data graphically:

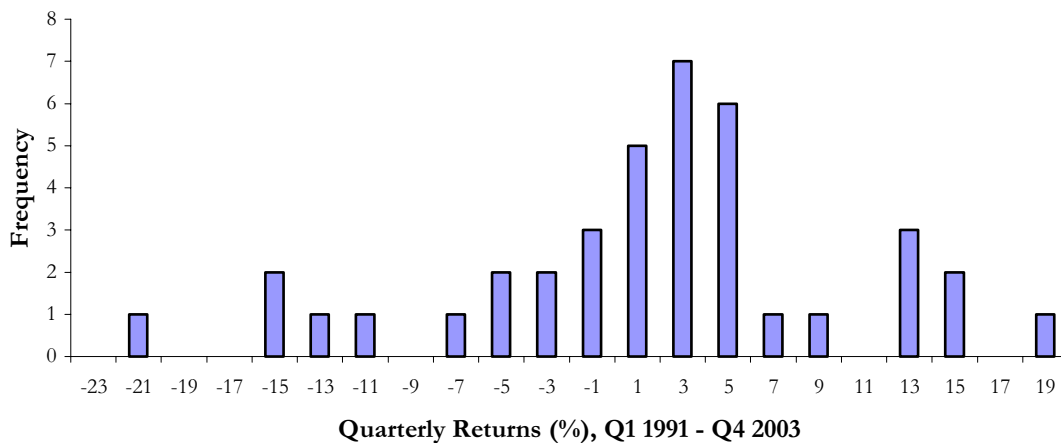
- frequency bar charts;
- relative frequency bar charts;
- cumulative frequency distributions or ogives;
- histograms.

II.B.2.3.1 The Frequency Bar Chart

To illustrate the frequency distribution with a bar chart, the frequency of observation is set on the vertical axis and the class interval (of the returns) is set on the horizontal axis. The frequencies are then plotted (using a ‘Column Chart’ in Excel) as in Figure II.B.3 and in the [Excel file](#). The height of each bar represents the frequency of observation within the corresponding class interval. Note that the labels on the category (x) axis have been adjusted so that the first bin, which represents returns from -24% to -22% , has been labelled with its midpoint, -23% .

Figure II.B.3

Frequency Distribution, MSCI World Equity Index



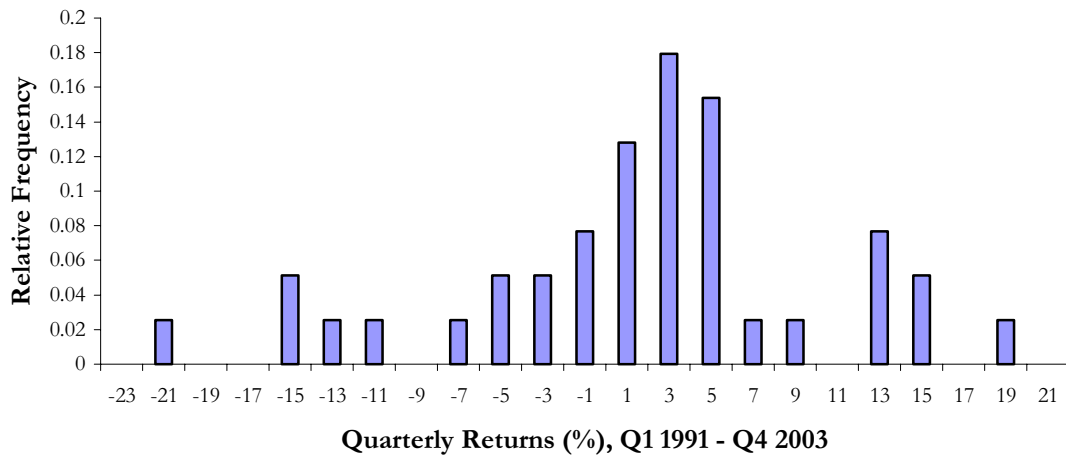
² Philosophically it is not necessary to consider this since no measure can be exact. We can always conceive of a more accurate measuring device, but all it will be able to tell us is whether that which we are measuring is, for instance, just above $+2$ or just below it.

II.B.2.3.2 The Relative Frequency Distribution

To derive the relative frequency of a given group of data, the frequency in that class interval must be divided by the total number of observations. The relative frequency distribution can then be plotted in a similar manner to the frequency distribution, producing a chart as shown in Figure II.B.4.

Figure II.B.4

Relative Frequency Distribution, MSCI World Equity Index

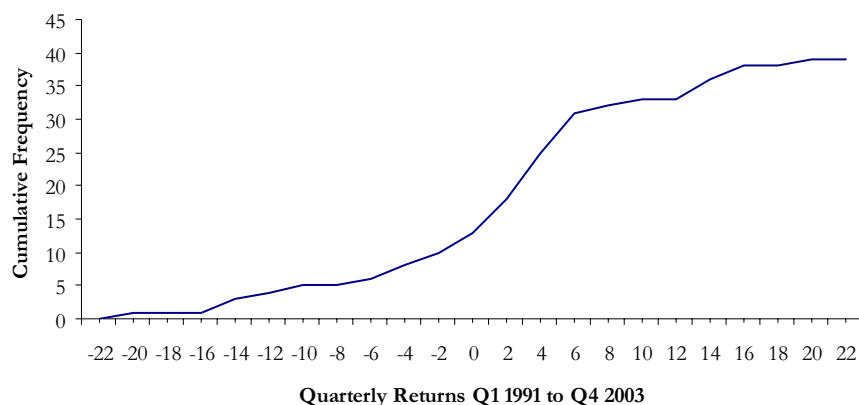


II.B.2.3.3 The Cumulative Frequency Distribution

To construct a cumulative frequency distribution, also known as an ‘ogive’, cumulative frequencies are plotted on the vertical axis and the class intervals are plotted along the horizontal axis. The graph is constructed by plotting the data in ascending order, and in the case of grouped data, plotting against the upper value of the class interval. This is illustrated in Figure II.B.5 (using a ‘Line Chart’ in Excel). The figure can be interpreted as follows: 13 observations from the sample have returns less than or equal to zero, 18 observations from the sample have returns less than or equal to 2%, and so forth. Note that for this graph the category axis is now continuous, with *points* being labelled, rather than intervals.

Figure II.B.5

Cumulative Frequency, MSCI World Equity Index



II.B.2.3.4 The Histogram

Data counts are, of course, discrete. But if they originate from a continuous set of possibilities then we should reflect that in our representation of the situation. We do this by creating a histogram, in which relative frequency (i.e. probability) is represented not by height, but by area. (See also Section II.E.2.2, where the link is made between histograms and probability density functions.)

The histogram in Figure II.B.6 is constructed so that the area of each column represents the corresponding relative frequency. It is easier to achieve this if the class intervals are all of equal width, though there are often occasions when this is not possible. The height of a bar is calculated by dividing the relative frequency that it represents by its width.

Again, the category axis is continuous, and points are labelled, not intervals. The frequency axis now represents relative frequency divided by width, so it is a *density* – relative frequency density. Here 9% of the observations in the sample are greater than or equal to 2% and less than 4%. In Section II.E.2.2 you will see how this is idealised to *probability density*.

By changing the width of the intervals different visualisations of the data may be constructed. In Figure II.B.7 bins of twice the width (4%) are used. So, 15% of observations in the sample are greater than or equal to 0% and are less than 4%.

Figure II.B.6

Histogram, MSCI World Equity Index

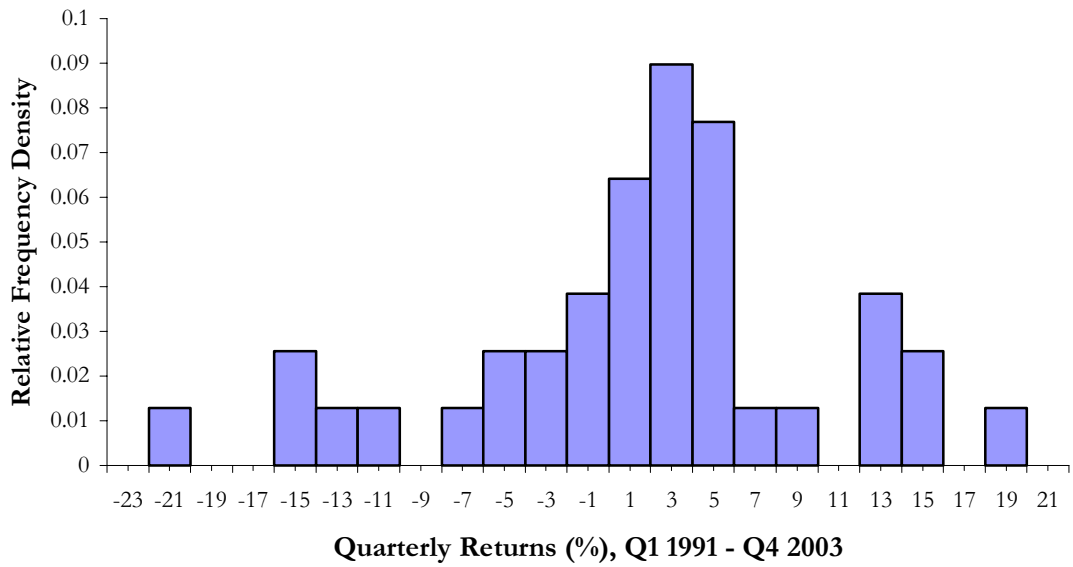
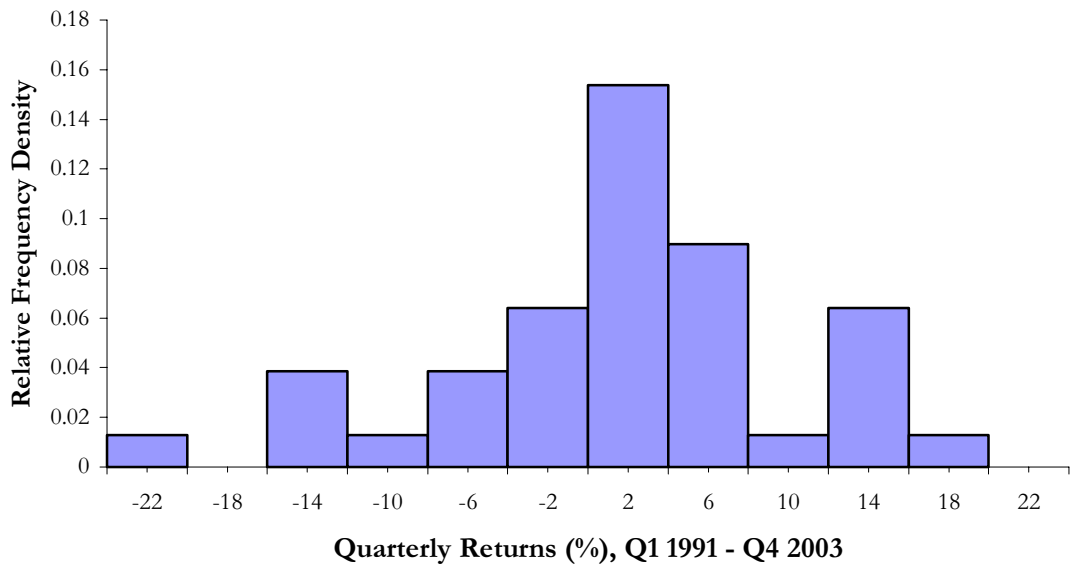


Figure II.B.7

Histogram, MSCI World Equity Index



II.B.3 The Moments of a Distribution

Moments capture qualities of the distribution of the data about an origin. The k th moment about an origin a is given by

$$\frac{\sum_{i=1}^n (x_i - a)^k}{n}. \quad (\text{II.B.1})$$

Thus to calculate the k th moment of a distribution about a point a , the deviation of each data point from a is raised to the power k . The results are then summed and the total divided by the total number of data points to give an average.

- If $a = 0$ and $k = 1$ we have what will be shown later to be the *arithmetic mean*. So the *mean* is sometimes said to be the *first moment about zero*.
- If a is the arithmetic mean and $k = 2$ we have the *second moment about the mean*. This is known as the *variance*, which is a measure of dispersion.
- If a is the mean and $k = 3$ we have the *third moment about the mean* which is a measure of *skewness*.
- If a is the mean and $k = 4$ we have the *fourth moment about the mean*, which measures *peakedness*.

The arithmetic mean and the variance are particularly important when data are *normally* distributed,³ which is often assumed to be the case for financial data. If data are normally distributed the characteristics of the distribution are completely specified by its arithmetic mean and its variance. The moments relating to skewness and kurtosis are often referred to as the *higher moments*.

II.B.4 Measures of Location or Central Tendency – Averages

There are several ‘averages’, each giving a measure of the location of the data. We will cover the arithmetic and geometric means, the median and the mode.

II.B.4.1 The Arithmetic Mean

The arithmetic mean is the most widely used measure of location and is what is usually considered to be *the* average. The arithmetic mean is calculated by summing all the individual observations and dividing by the number of those observations. For example, assume that we wish to calculate the arithmetic mean quarterly return of an asset over five consecutive quarters.

³ See Section II.E.4.4. Whether or not the data is normally distributed is an empirical issue that we will touch on in Chapter II.F.

Assume that the returns are as follows: 5%, 2%, -3%, 4% and 2%. The arithmetic mean is calculated by summing the five observations and dividing by the number of observations. Symbolically:

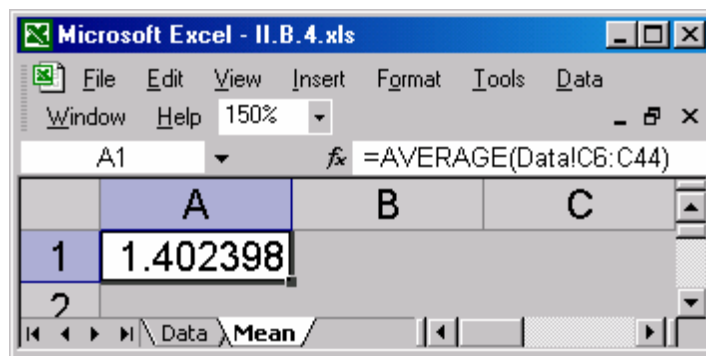
$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} . \quad (\text{II.B.2})$$

The Σ (the Greek capital letter sigma) is the *summation operator*, i is the index and \bar{x} is the symbol for the mean of the data items $x_1, x_2, x_3, \dots, x_n$. In our example there are five data items,

$$\sum_{i=1}^5 x_i = 5 + 2 + (-3) + 4 + 2 = 10, \text{ and so } \bar{x} = \frac{10}{4} = 2.5 .$$

Now let us apply this to the MSCI equity returns data. The appropriate Excel function is (perhaps unfortunately) =AVERAGE(...). Figure II.B.8 shows that the arithmetic mean of the continuously compounded monthly return to this index over the period in question is 1.4024%.

Figure II.B.8



Note that the arithmetic mean in statistics (\bar{x}) corresponds to the *expected value* in probability (μ). So \bar{x} is an estimate for μ . Note also that the arithmetic mean is particularly susceptible to extreme values. Extremely high values increase the mean above what is actually representative of the point of ‘central tendency’ of the data. Extremely low values have the opposite effect.

II.B.4.2 The Geometric Mean

An alternative average is the *geometric mean*. This is particularly applicable when the interest on an investment is discretely compounded and is reinvested.⁴ To illustrate this, assume that a hedge fund generated the following annual rates of return over five years: +10%, +20%, +15%, –30%, +20%. The arithmetic mean is $+35/5 = 7\%$. However, \$100 invested in year one would grow over the five years as follows: $\$100 \times 1.10 \times 1.20 \times 1.15 \times 0.70 \times 1.20 = 127.51$. Thus the actual growth over the whole of the five-year period is only 27.51%.

Clearly we need a single measure of a growth rate which if repeated n times will transform the opening value into the terminal value. We first compute the geometric mean of the compounding factors:

$$\bar{y}_g = \sqrt[n]{y_1 \times y_2 \times y_3 \times \dots \times y_n} \quad (\text{II.B.3})$$

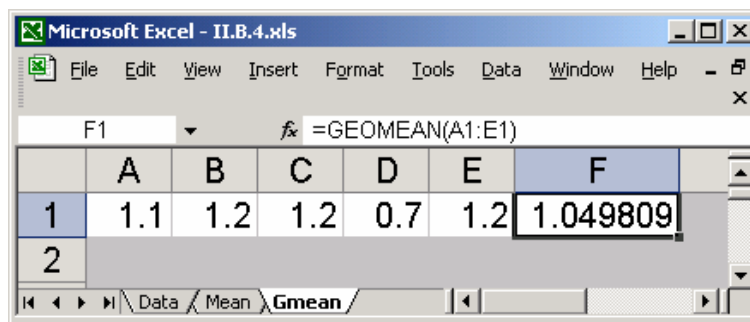
where the $y_i = (1 + r_i)$ and r_i is the rate of return for the i th period expressed in decimals, for example $10\% = 0.1$. The required rate, \bar{r}_g , is then given by $\bar{r}_g = \bar{y}_g - 1$.

Using the above data, the geometric mean rate of growth for the five years in question will be

$$\bar{r}_g = \sqrt[5]{1.1 \times 1.2 \times 1.15 \times 0.7 \times 1.2} - 1 = 1.0498 - 1 = 4.98\%,$$

as shown in Figure II.B.9.

Figure II.B.9



⁴ Note the interplay between the geometric mean for discretely compounded returns, and the arithmetic mean for continuously compounded returns. If five consecutive prices of an asset are p_1, p_2, p_3, p_4 and p_5 , then the geometric mean of the compounding factors is $\sqrt[4]{\frac{p_2}{p_1} \times \frac{p_3}{p_2} \times \frac{p_4}{p_3} \times \frac{p_5}{p_4}} = \sqrt[4]{\frac{p_5}{p_1}}$. The arithmetic mean

of the continuously compounded returns is $\frac{\ln\left(\frac{p_2}{p_1}\right) + \ln\left(\frac{p_3}{p_2}\right) + \ln\left(\frac{p_4}{p_3}\right) + \ln\left(\frac{p_5}{p_4}\right)}{4} = \frac{\ln\left(\frac{p_5}{p_1}\right)}{4} = \ln\left(\sqrt[4]{\frac{p_5}{p_1}}\right)$ so

the arithmetic mean of the continuously compounded returns is log of the geometric mean of discretely compounded returns.

To find the annualised geometric mean return of an asset over a given number of compounding periods we have to use the starting value and the end value:

$$\bar{r}_g = \left(\frac{v_n}{v_0} \right)^{1/n} - 1, \quad (\text{II.B.4})$$

where v_n is the terminal price (or value) of the asset after n compounding periods and v_0 is the initial price (or value) of the asset.

Consider the data for the MSCI equity index. The level of that index on 31 March 1993 was 599.74. The value on 31 December 2003, $9\frac{3}{4}$ years later, was 1036.23. The geometric mean quarterly return over that period was:

$$\bar{r}_g = \left(\frac{1036.23}{599.74} \right)^{1/39} - 1 = 0.014121$$

or approximately 1.41% per quarter. Annualising gives $1.014121^4 - 1 = 0.0577$, or 5.77%.

II.B.4.3 The Median and the Mode

To find the *median* we must first arrange the data in order of size. If there is an odd number of data items then the median is the middle observation. If there is an even number of data items then the median is the mean of the middle two items. The median is generally held to be an extremely useful measure of location – it is representative and robust. It is not affected by outliers. Its disadvantage is that, perhaps surprisingly, it is difficult to compute – arranging data in order is a more exacting computational task than summing the items. Furthermore, the process leads into an area of statistics known as *order statistics* in which the convenient techniques of calculus cannot be applied. This, together with the importance of the mean as one of the parameters of the normal distribution, means that it is much less used than the mean. Nevertheless it *is* used, and you will see a specific use for it in Section II.B.5.5.

The mode is the most commonly occurring item in a data set. It is perhaps most useful in a descriptive context when *bimodal* distributions arise. For instance, over a period of time the interest rates charged to a capped mortgage might have one frequency peak at a value below the cap, and one peak at the cap.

II.B.5 Measures of Dispersion

As well as having a measure of central location, we need to know how the data are dispersed around that central location. We will see later in this course that we use measures of dispersion to measure the financial risk in assets and portfolios of assets. The following are the measures of dispersion that we shall study:

- the variance;
- the standard deviation;
- the negative semi-variance (other downside risk metrics are covered in Section I.A.1.7.4).

II.B.5.1 Variance

The variance is widely used in finance as a measure of uncertainty, and has the attractive property of being additive for independent variables: $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ if X and Y are uncorrelated (see Section II.B.6). The standard deviation has a similar use and is employed both to measure risk in financial assets and as the measure of volatility in pricing options. However, because the standard deviation is the square root of the variance, it is not additive.

If the individual observations are closely clustered around the mean, the differences between each individual observation, x_i , and the mean, \bar{x} , will be small. If the individual observations are widely dispersed, the difference between each x_i and \bar{x} will be large.

It may be thought that by summing $x_i - \bar{x}$ we get a measure of dispersion. However, that is not the case since $\sum(x_i - \bar{x})$ is always zero. To overcome this problem we square the individual deviations, and sum the squared figures. Dividing the result by $n - 1$, the number of observations less one (see below), gives the sample variance. The symbol used for this is s^2 , so:

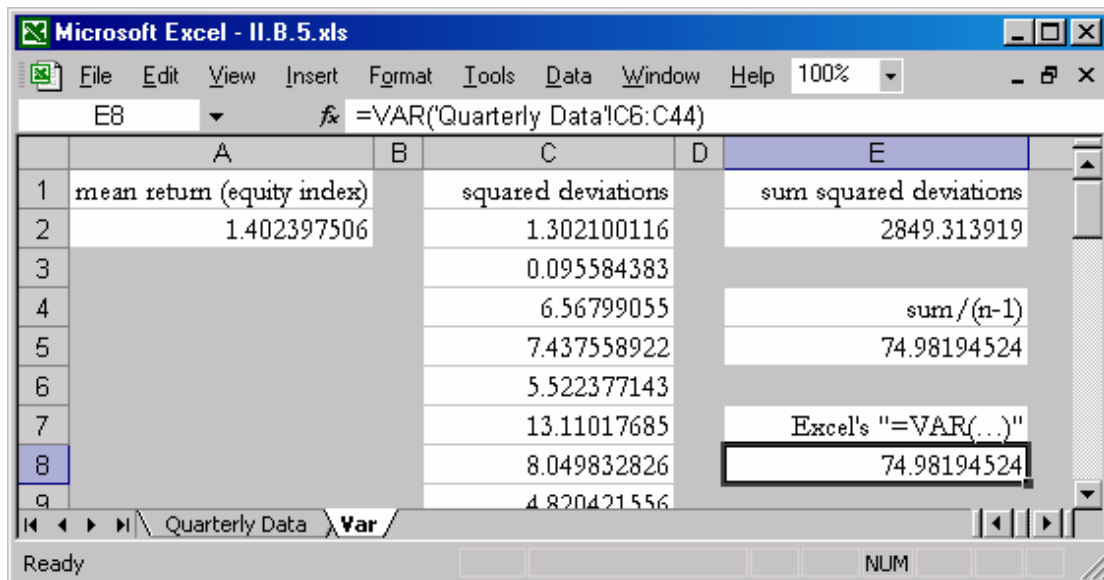
$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}. \quad (\text{II.B.5})$$

Note that if the variance is derived from the whole population of data then $\bar{x} = \mu$, the divisor in equation (II.B.5.1) would be n , and σ^2 would be the appropriate notation for the variance – we would be dealing with *parameters* and not with *statistics* (see Section II.B.3 above). However, when the data only represent a *sample* of the population data, as it usually does in empirical research in finance, we must divide by $n - 1$. This exactly compensates for having to use \bar{x} in the calculation instead of μ , thus avoiding any bias in the result. That this compensation is exact can be proved in half a page of algebra, but we will not inflict that on you. Rather we refer you to the concept of *degrees of freedom*. This is commonly used, though it represents only a partial explanation. The number of degrees of freedom is the number of observations minus the number

of parameters estimated from the data. In the case of the variance, the mean is estimated from the same data. Thus, when using that estimate in computing the standard deviation, the number of degrees of freedom is $n - 1$. The effect of replacing n by the number of degrees of freedom is greatest with small samples: put another way, the relative difference between n and $n - 1$ is smaller when n is larger. With financial data we are often operating with large samples that have very small means, so sometimes we estimate variance as $\sum x_i^2 / n$, although s^2 would be correct.

To illustrate the calculation of the variance again refer to Figure II.B.10 which shows the first few lines of a detailed computation of variance, together with the result of applying the Excel spreadsheet function =VAR(...). The variance is computed as 74.9819. This figure represents squared percentages, which does not lend itself to an intuitive interpretation! Nevertheless, variance is at the root of much of our work in risk control.

Figure II.B.10



II.B.5.2 Standard Deviation

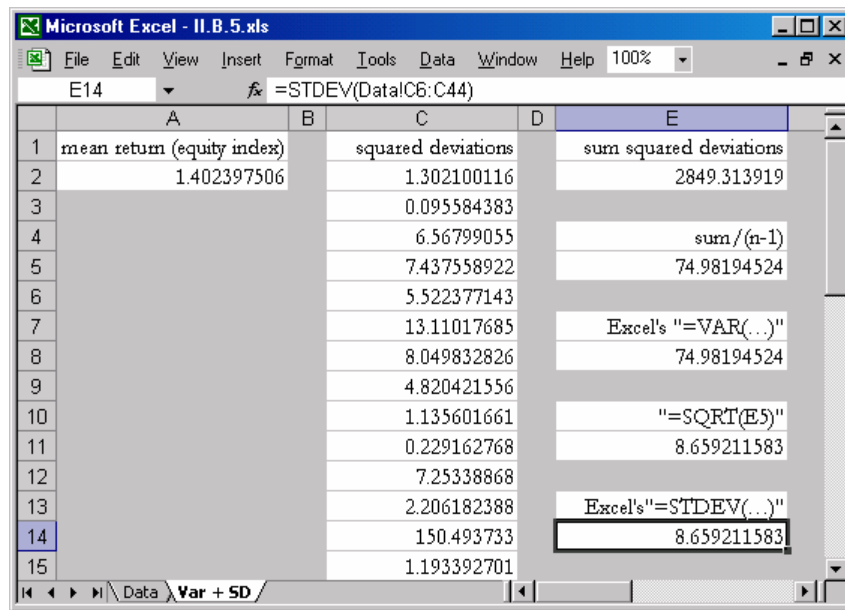
The variance is in squared units of the underlying data, which makes interpretation rather difficult. This problem is overcome by taking the square root of the variance, giving the *standard deviation*. The usual notation, unsurprisingly, is to use s for the statistic and σ for the parameter. The former is for the sample standard deviation, using an $(n - 1)$ divisor; the latter is for the population standard deviation.

Thus the formula for the standard deviation as

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}} \quad (\text{II.B.6})$$

Referring to the calculation of the variance from the data of the MSCI World Equity Index returns, the square root of 74.9819 is 8.66. Thus the standard deviation is 8.66% (see Figure II.B.11).

Figure II.B.11



II.B.5.3 Case Study: Calculating Historical Volatility from Returns Data

You will learn in other chapters of the *PRM Handbook* that volatility is a very important parameter in the pricing of financial options. Historical volatility is often used as a basis for forecasting future volatility. Volatility is almost always quoted in annual terms, so historical volatility is the annualised standard deviation of the continuously compounded returns to the underlying asset. We know how to calculate the standard deviation, so we now only have to understand the process of *annualisation*.

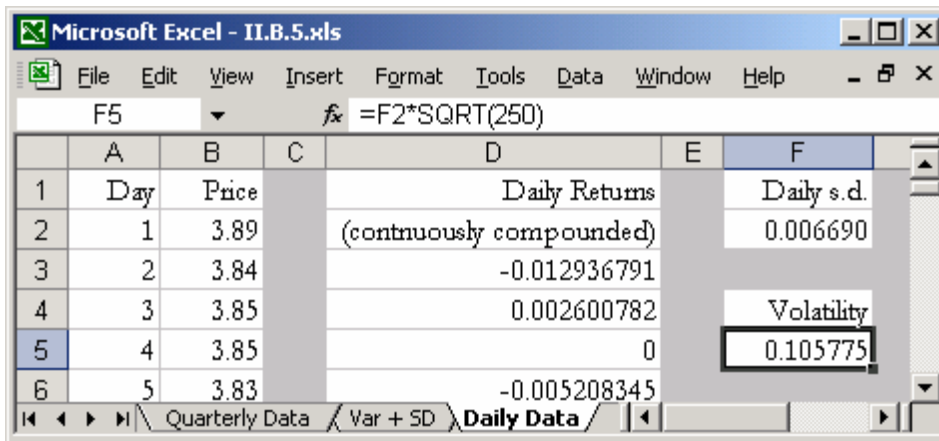
The annualisation process assumes that individual returns observations are independent and identically distributed (i.i.d.). ‘Independent’ means that the correlation coefficient between successive observations is zero. ‘Identically distributed’ means that they are assumed to come from the same probability distribution – in this case a normal distribution (see Section II.E.4.4). Under these assumptions the variance of the returns is a linear function of the time period being analysed. For example, the variance of one-year returns could be calculated as 12 times the

variance of one-month returns, or 52 times the variance of one-week returns or 250 times⁵ the variance of daily returns.

However, as the standard deviation is the square root of a variance it is not additive. To annualise a standard deviation we must multiply the calculated number by the *square root* of the number of observation periods in a year. Thus if daily data are available we first calculate the standard deviation of daily returns. Then to arrive at the annualized standard deviation, the volatility, we multiply the daily figure by the square root of the number of trading days in a year. Similarly, if weekly data are available we first calculate the standard deviation of weekly returns. Then to arrive at the annualized standard deviation, the volatility, we multiply the weekly figure by the square root of 52, there being 52 weeks in a year. Or for monthly data, the standard deviation of monthly returns is multiplied by the square root of 12 to get the volatility. For instance, using the MSCI equity index data, the standard deviation of quarterly returns is 8.66%. Therefore the annualised standard deviation, or volatility, is $8.66 \times \sqrt{4} = 8.66 \times 2 = 17.32\%$.

The Excel workbook [II.B.5.xls](#) contains 57 daily observations of the prices of an asset. The standard deviation of daily returns is 0.006690 (0.6690%) and therefore the volatility is $0.006690 \times \sqrt{250} = 0.105775 \approx 10.6\%$ (see Figure II.B.12).

Figure II.B.12



Note that we have calculated the annualised volatility without using one year of data. We have sampled from 57 daily returns observations in order to calculate the standard deviation of daily returns, and then annualised the result. How valid is the sampling over just 57, or any other number of days, will depend upon how representative were those days.

⁵ This assumes that there are 250 trading days in a year, allowing for weekends and public holidays. Actually the number will be between 250 and 260, depending on the market and the particular year.

II.B.5.4 The Negative Semi-variance and Negative Semi-deviation

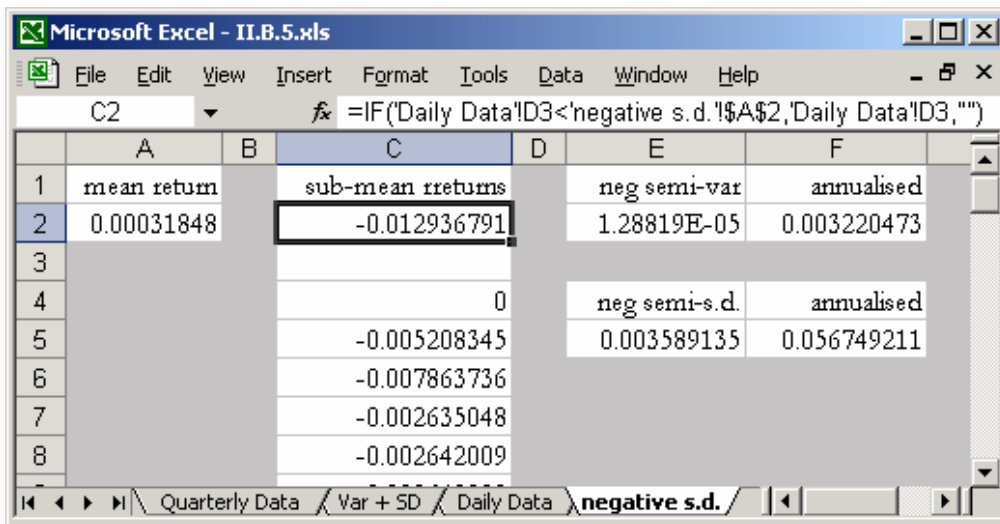
The *negative semi-variance* is similar to the variance but only the negative deviations from the mean are used. The measure of negative semi-variance is sometimes suggested as an appropriate measure of risk in financial theory when the returns are not symmetrical (see Chapter I.A.1). It is, after all, downside risk which matters! It is calculated as follows:

$$\bar{s}^2 = \frac{\sum (r_i - \bar{r})^2}{n - 1}, \quad (\text{II.B.7})$$

where r_i are portfolio returns which are less than the mean return, n is the total number of these negative returns, \bar{r} is the mean return for the period including positive and negative returns, and \bar{s}^2 is the semi-variance. The square root of the negative semi-variance is the negative semi-deviation, \bar{s} .

Figure II.B.13 shows a computation of these quantities for the daily data in [II.B.5.xls](#). The logical IF operator has been used to select those returns which are less than the mean. Those which are larger have been replaced by an empty space, indicated by the double quotation, "", in the third argument position of the IF function in the highlighted cell. Excel ignores empty spaces in the subsequent variance and standard deviation computations.

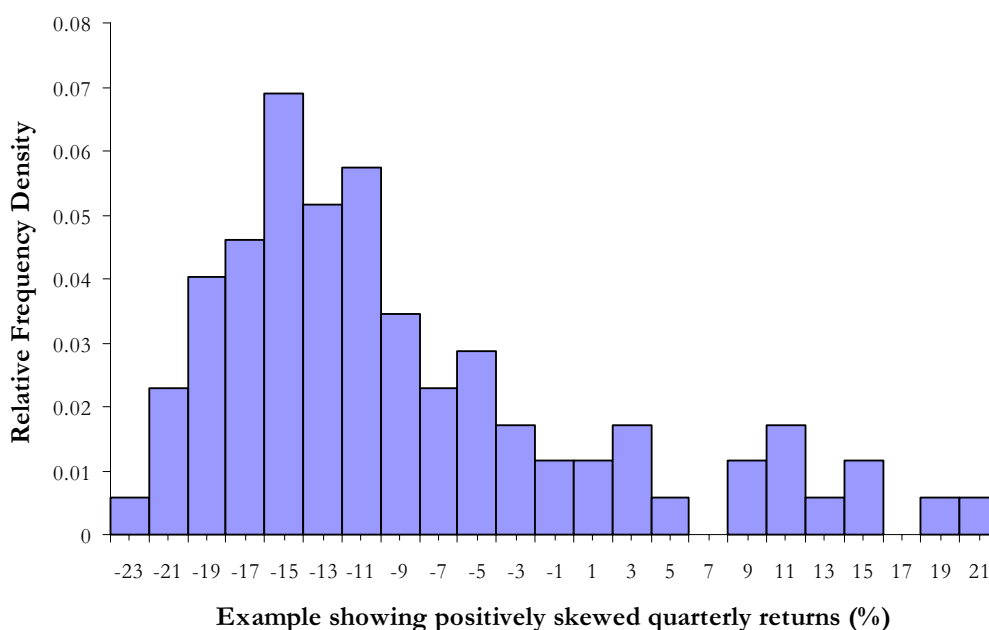
Figure II.B.13



II.B.5.5 Skewness

It is important to consider if there is any bias in the dispersion of the data. If the data is symmetric then the variance (standard deviation) statistic captures its dispersion. But if the data is not symmetric that is not the case. Lack of symmetry (or bias) is measured by the skewness. In the case of positive skewness the distribution has a long tail to the right (as in Figure II.B.14). Negative skewness results in a longer tail to the left.

Figure II.B.14



The discrete compounding of periodic returns will give rise to positive skewness. For instance, consider the terminal value of an asset that showed positive returns of 8% p.a. for two annual periods. For an original investment of 100, the terminal value would be

$$100 \times 1.08^2 = 116.64.$$

Now if the returns were -8% in each year the terminal value would be

$$100 \times 0.92^2 = 84.64.$$

Thus compounding the positive returns gives a gain of 16.64 over the two-year period, whereas the two negative returns only lose 15.36. This is one important reason why it is preferable to use continuously compounded or log returns, which are more likely to exhibit symmetry.

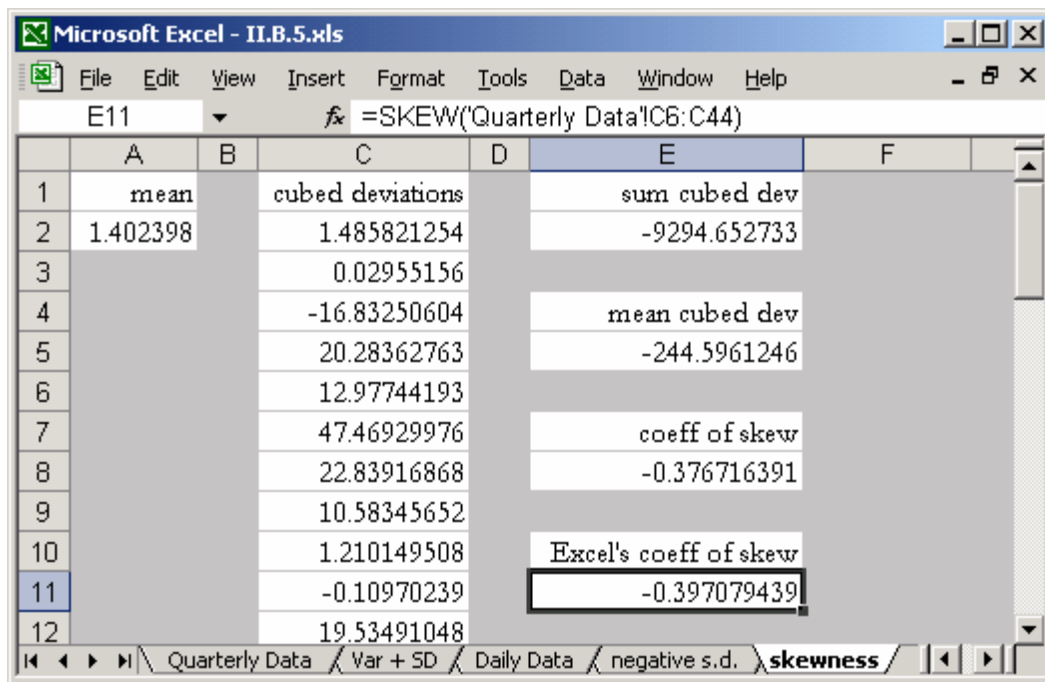
Recall from Section II.B.4.3 that the median is not affected by the size of extreme values, whereas the mean is. Thus positive skewness is indicated if the mean is significantly larger than the median, and negative skewness is indicated if the mean is significantly less than the median.

However, differences between the mean and the median only indicate the presence of skewness. What is needed is a measure of it. An appropriate measure is the *coefficient of skewness* (see Section II.B.3). This is derived by calculating the third moment about the mean, which is then standardised by dividing by the cube of the standard deviation:

$$\frac{\left[\frac{\sum (x - \bar{x})^3}{n - 1} \right]}{\left(\sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \right)^3} \quad (\text{II.B.8})$$

Figure II.B.15 shows a computation of skewness for the MSCI World Equity Index returns. The numerator of equation (II.B.8) evaluates to -244.5961. Dividing by the cube of the standard deviation gives a moment coefficient of skewness of -0.3767.

Figure II.B.15



If the distribution of the equity index returns were symmetrical, the coefficient of skewness would be zero. As this calculation shows a negative number, the index returns data is negatively skewed.⁶ This implies that large negative returns are more likely than large positive returns of the same size.

⁶ This does not contradict our earlier assumption that we might expect to see a positive skew when looking at returns. There we were simply looking at discretely compounded returns. Here we have used continuously compounded returns.

Note that there is often an important *leverage* effect underlying skewed returns distributions. For instance, a large stock price fall makes the company more highly leveraged and hence more risky. Hence it becomes more volatile and likely to have further price falls. The opposite is the case for commodities, where in that context a price *rise* is bad news.

This raises the question of whether our observed coefficient is *significantly* different from zero. We shall examine such issues in Chapter II.F, where we deal with statistical hypothesis testing. Note that Excel’s coefficient of skewness, shown in the screen shot, differs slightly from the moment definition which we have used. It is the moment definition multiplied by $n/(n - 2)$. (In the example $n = 39$.) This is to facilitate hypothesis testing.

II.B.5.6 Kurtosis

Whereas skewness indicates the degree of symmetry in the frequency distribution, kurtosis indicates the ‘peakedness’ of that distribution. Distributions that are more peaked than the normal distribution (which will be discussed in detail later) are referred to as *leptokurtic*.⁷ Distributions that are less peaked (flatter) are referred to as *platykurtic*, and those distributions that resemble a normal distribution are referred to as *mesokurtic*.

Leptokurtic distributions have higher peaks than normal distributions, but also have more observations in the tails of the distribution and are said to be *heavy-tailed*. Leptokurtic distributions are often found in asset returns, for instance when there are periodic jumps in asset prices. Markets where there is discontinuous trading, such as security markets that close overnight or at weekends, are more likely to exhibit jumps in asset prices. The reason is that information that has an influence on asset prices but is published when the markets are closed will have an impact on prices when the market reopens, thus causing a jump between the previous closing price and the opening price. This jump in price, which is most noticeable in daily or weekly data, will result in higher frequencies of large negative or positive returns than would be expected if the markets were to trade continuously.

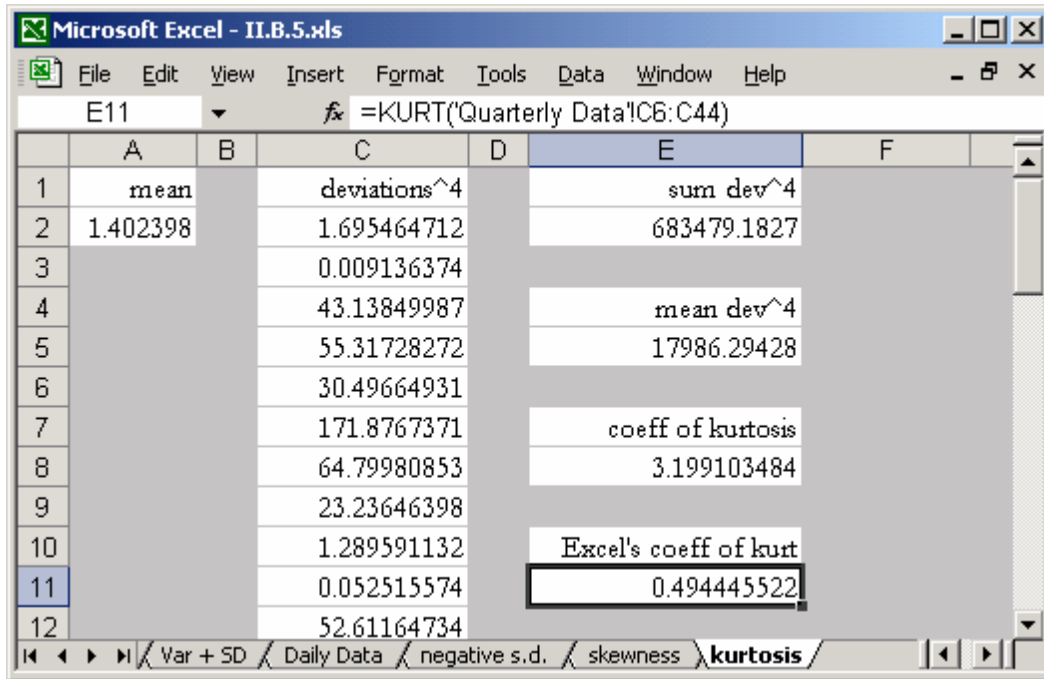
The *coefficient of kurtosis* is derived by standardising the fourth moment about the mean by dividing by the standard deviation raised to the fourth power:

$$\frac{\frac{\sum(x - \bar{x})^4}{n - 1}}{\left(\sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}\right)^4} \quad \text{(II.B.9)}$$

⁷ Literally meaning ‘thin arches’ in Greek.

Figure II.B.16 shows the computation of kurtosis for the MSCI World Equity Index returns. The numerator of equation (II.B.9) evaluates to 17,986.29. Dividing by the fourth power of the standard deviation gives a moment coefficient of kurtosis of 3.199.

Figure II.B.16



If the data were normally distributed (i.e. mesokurtic) the moment coefficient of kurtosis would be 3.0. The computed index value is greater than 3, indicating that the returns data in question are more peaked than a normal distribution and therefore are leptokurtic. If the data were platykurtic, the moment coefficient of kurtosis would be less than 3.0.

Again the question is whether our observed coefficient is *significantly* different from 3. Excel's coefficient of kurtosis, shown in Figure II.B.15, is different. It is given by

$$\frac{n(n+1)}{(n-2)(n-3)} \times \text{moment coefficient} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Note that the Excel function 'KURT' is a measure of *excess* kurtosis; a normal distribution will have excess kurtosis approximately equal to 0 (rather than 3).

II.B.6 Bivariate Data

There are many occasions when our observations are not just of a single variable (univariate statistics), but of two jointly varying variables (bivariate statistics) or more than two jointly varying

variables (multivariate statistics). In this section we look at bivariate data, and consider *measures of association*. The material in this section is closely related to other parts of the *PRM Handbook* and to Section II.E.3 in particular, where we discuss the joint probability distribution that can be used to model bivariate data.

Earlier it was explained that the variance indicated how observations of a single variable are distributed around the mean observation. In this section we will introduce the concept of the *covariance*, which indicates how the observations of two variables behave in relation to each other. Later in this section we will go on to calculate the *correlation coefficient*, which is a more convenient measure of the linear association between two variables. In many areas of risk management it is necessary to know how the returns of security X behave over time in relation to the returns of security Y , for all possible pairs (X, Y) . We have two related measures of co-variability – *covariance* and *correlation*. Correlation is just a standardised form of covariance, but we shall need both measures. These are fundamental concepts to understand when estimating portfolio risk (see Sections I.A.2.1, II.D.2.1, and many others) and constructing diversified portfolios (see, for example, Sections II.D.2.2 and I.A.3.1).

II.B.6.1 Covariance

If the returns to security 1 generally rise (fall) at the same time that the returns of security 2 rise (fall), the covariance will be positive. However, if generally the returns of security 1 rise while the returns of security 2 fall, the covariance will be negative. If there is no pattern to the relationship returns, that is, if the two security returns are totally independent, then the covariance will be zero.

The formula for the covariance is

$$\text{Cov}_{XY} = s_{XY} = \frac{\sum_i ((x_i - \bar{x}) \times (y_i - \bar{y}))}{n - 1}. \quad (\text{II.B.10})$$

Note that under this definition $\text{Cov}_{XX} = \text{Var}_X$.

The size of the covariance depends on the magnitude of the observations x_i and y_i . Therefore a larger covariance could be more to do with having high values to the observations than with a closer association between the variables. We will return to this issue later in our discussion of the correlation coefficient.

Figure II.B.17 shows a computation of covariance of the returns to the MSCI World Equity Index and the returns to the MSCI Sovereign Government Bond Index. The covariance is computed to be -5.2178 .

Figure II.B.17

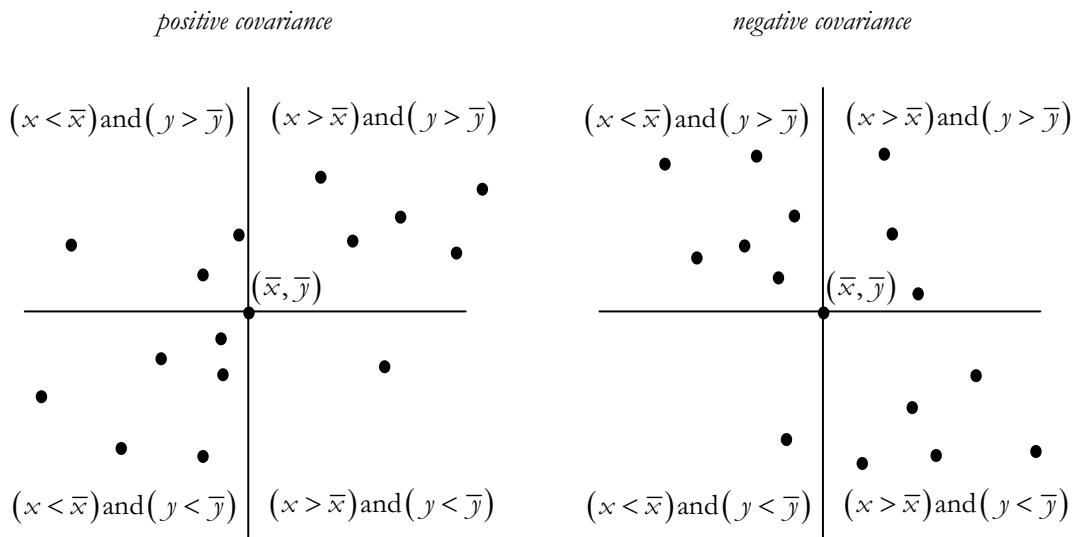
	A	B	C	D	E	F	G
1	equity mean		equity deviations	bond deviations	cross product		computed cov
2	1.402398		1.141096	-1.080272	-1.232694		-5.217756
3			0.309167	-0.591681	-0.182928		
4	bond mean		-2.562809	-1.191175	3.052755		Excel's cov
5	1.673737		2.727189	8.747739	23.856735		-5.083967
6			2.349974	3.551537	8.346020		

Yet again note that the Excel version is slightly different from the computed value. This is because Excel’s COVAR function uses a divisor of n instead of $n - 1$. This is not consistent with its correlation function, CORREL, where a divisor of $n - 1$ is used. This can be corrected by multiplying the Excel covariance by $n/(n - 1)$, but this will not make much difference for large data sets.

We see that in this example, assets 1 and 2 have a negative covariance between their returns. To understand how the sign of the covariance develops refer to Figure II.B.18, which has been divided into four by two perpendicular lines through the point (\bar{x}, \bar{y}) . For points in the top left-hand quarter, the values of y are greater than \bar{y} , thus $y - \bar{y}$ is positive but values of x will be below \bar{x} , and so $x - \bar{x}$ will be negative. Thus the product is negative, so the contribution of observations in the top left-hand quarter will be negative. By analogous reasoning, the contribution of any observations in the top right-hand corner will be positive, as will the contribution of those in the bottom left-hand corner. The contribution of those observations in the bottom right-hand corner will be negative.

Thus data that are predominately located in the bottom left and top right quarters will have a positive covariance, whilst data where the observations predominate in the top left and bottom right quarters will have a negative covariance. If the data points are scattered evenly over the four quarters, the covariance will be close to zero.

Figure II.B.18



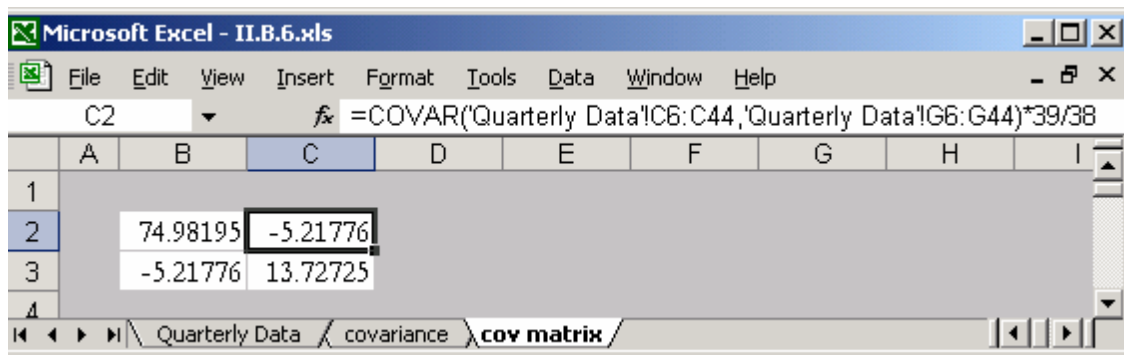
II.B.6.2 The Covariance Matrix

The covariances between several pairs of variables are usually displayed in matrix format, as in the following structure for three variables.

$$\begin{bmatrix} \text{Var}_X & \text{Cov}_{XY} & \text{Cov}_{XZ} \\ \text{Cov}_{YX} & \text{Var}_Y & \text{Cov}_{YZ} \\ \text{Cov}_{ZX} & \text{Cov}_{ZY} & \text{Var}_Z \end{bmatrix},$$

This shows the variances and covariances between all possible pairs from a group of three assets. Covariance matrices are of fundamental importance in portfolio optimisation and play a central role in many other chapters of the *PRM Handbook* – see Chapters I.A.2, I.A.3, III.A.3 as well as Section II.D.2, for instance. Figure II.B.19 shows the covariance matrix for the MSCI indices.

Figure II.B.19



II.B.6.3 The Correlation Coefficient

You will remember that variance, whilst being much used, is not easily interpretable. Taking the square root leads to the standard deviation which is commensurate with the data, and which is easy to interpret.

Covariance suffers similarly from problems of interpretation in that it depends on the units in which each variable is measured. In most cases we will be looking at returns, so both variables will be unit-free, but this may not always be the case. We need a unit-free measure of association – one that will not be affected by arbitrary decisions regarding units of measure (e.g. pounds or pence). We achieve this by dividing the covariance by the product of the standard deviations. This gives the correlation coefficient, r (or ρ for the corresponding population parameter):

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} . \quad (\text{II.B.11})$$

Note that we often have to use equation (II.B.11) in the form

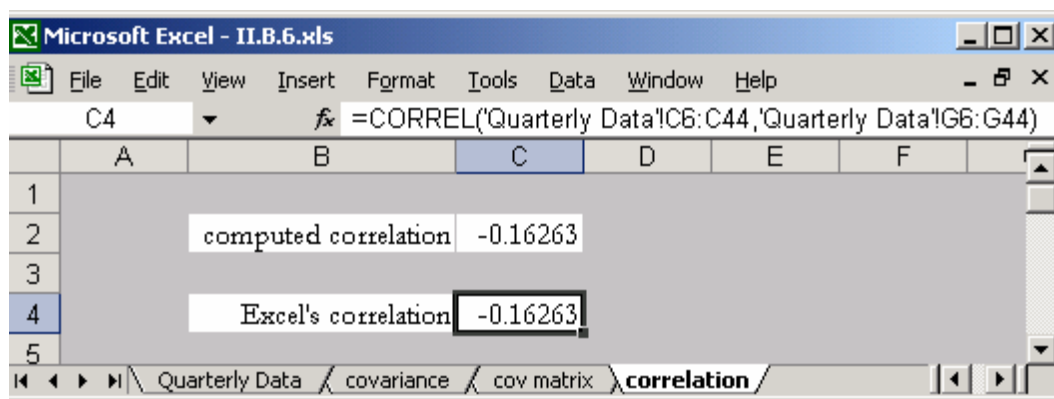
$$s_{XY} = r_{XY} s_X s_Y .$$

Note also that in these notes we use the notation r_i for the returns to asset i , whilst r with two subscripts denotes a sample correlation coefficient.

The correlation coefficient is a unit-free measure of the strength and the direction of a linear relationship between two variables. The values of the correlation coefficient range from -1 for a perfectly negative relationship, through zero where the two variables are linearly independent of each other, to $+1$ for a perfectly positive relationship between the variables.

Figure II.B.20 shows the correlation between the MSCI indices.

Figure II.B.20

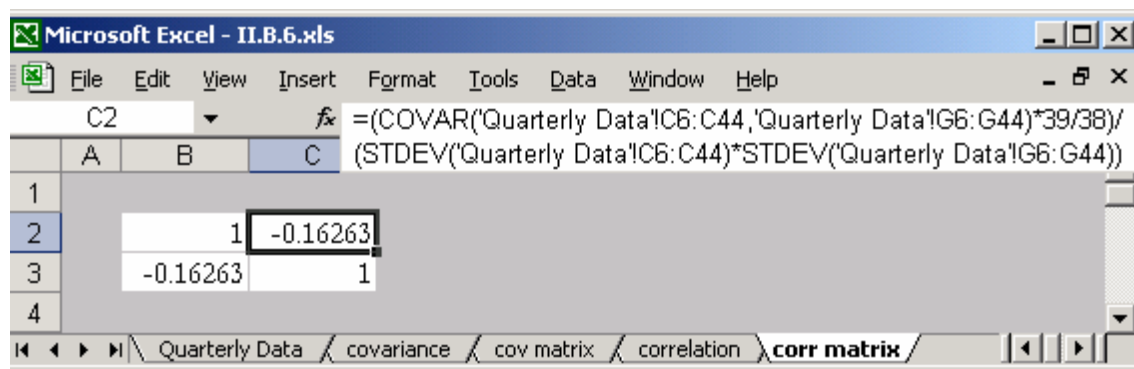


Irrespective of how positive or negative the correlation is, the correlation coefficient is only a measure of statistical association. There is no inference of causality in the statistic. By this we mean that there is no suggestion that a change in variable X causes a change in variable Y , or indeed vice versa. There may, for instance, be some underlying mechanism that is causing them both to vary together. Thus although the correlation coefficient may indicate how strong the linear association is between two variables, it cannot explain the changes in them. For such an explanation it is first necessary to develop a theory of the causal relationship from a priori reasoning, construct a model that reflects the hypothesised relationship and then test the model statistically. For this regression analysis is used. This is the subject of Chapter II.F.

II.B.6.4 The Correlation Matrix

Just as covariances are frequently displayed in matrices, so correlation coefficients are similarly displayed. Of course, the leading diagonal (top-left to bottom-right) elements of the correlation matrix will all be 1. Figure II.B.21 shows the correlation matrix for the MSCI indices.

Figure II.B.21



II.B.6.5 Case Study: Calculating the Volatility of a Portfolio

In this subsection we apply what we have learned about variances, standard deviations, covariances and correlations, as well as the annualisation of standard deviations. The objective is to calculate the volatility of a portfolio that has equal proportions of the MSCI Equity and Bond indices, in [II.B.1.xls](#).

The risk in a single asset is measured by the annualised standard deviation of the returns to that asset, and the risk of a portfolio is measured by the variance or standard deviation of the returns to that portfolio. Our first step must be to calculate the standard deviations of the individual assets.

To calculate the risk of a portfolio, we also need to know how the returns of pairs of assets fluctuate together. We need to know the covariances or alternatively the correlations, and we need to have this information for each pair of assets in the portfolio.

The portfolio risk, as measured by its variance, is calculated as the weighted sum of the covariances between each pair of assets in the portfolio, where each covariance is multiplied by the product of the weights of each of the respective assets in the pair, and where the variance of a particular asset is considered as the covariance of the asset with itself. An asset's weight is the proportion of the total value of the portfolio that is invested in that particular asset.

To demonstrate the calculation of portfolio risk, consider a portfolio of three securities 1, 2 and 3 with prices $S_1(t)$, $S_2(t)$ and $S_3(t)$ at time t . Consider a portfolio of n_1 units of security 1, n_2 units of 2 and n_3 units of 3. Suppose there are no dividends (if the security is an equity) or coupons (if the security is a bond), denote the value of the portfolio at time t by $p(t)$ and suppose there is no rebalancing over time. Then at any time $t \geq 0$,

$$p(t) = n_1 p_1(t) + n_2 p_2(t) + n_3 p_3(t).$$

Denote the portfolio return over the period from time 0 to time t by r_p . Then

$$(1 + r_p) = p(t)/p(0).$$

So we can also write

$$(1 + r_p) = \frac{n_1 p_1(t) + n_2 p_2(t) + n_3 p_3(t)}{p(0)} = \frac{n_1 p_1(0)}{p(0)} \frac{p_1(t)}{p_1(0)} + \frac{n_2 p_2(0)}{p(0)} \frac{p_2(t)}{p_2(0)} + \frac{n_3 p_3(0)}{p(0)} \frac{p_3(t)}{p_3(0)}$$

so letting

$$w_i = \frac{n_i p_i(0)}{p(0)} = \text{the proportion invested in asset } i,$$

we have

$$(1 + r_p) = w_1(1 + r_1) + w_2(1 + r_2) + w_3(1 + r_3) \tag{II.B.12}$$

From this it follows that⁸

$$\begin{aligned} \text{Var}(r_p) &= w_1^2 \text{Var}(r_1) + w_2^2 \text{Var}(r_2) + w_3^2 \text{Var}(r_3) \\ &\quad + 2w_1 w_2 \text{Cov}(r_1, r_2) + 2w_1 w_3 \text{Cov}(r_1, r_3) \\ &\quad + 2w_2 w_3 \text{Cov}(r_2, r_3). \end{aligned} \tag{II.B.13}$$

⁸ Result (II.B.13) is not actually proved, but more of the background will be covered in Section II.E.3.4.

The portfolio risk, as measured by its volatility, is annualised standard deviation, and the standard deviation is the square root of portfolio returns variance (II.B.13).

In Section II.D.2 we show how (II.B.13) can be conveniently expressed using matrix notation, which is useful when there are a large number of securities on the portfolio. But here we have just three. However, it is useful to recast (II.B.13) in terms of the correlation coefficient, rather than covariance, since that is frequently used as an alternative measure of association. The advantage of ranking pairs of assets by their correlation coefficients is that it provides a clear system for including assets that enhance the benefits of diversification, and excluding those that do not.

Applying this to a two-asset portfolio, and demonstrating the use of alternative notation, we have

$$s_p^2 = w_1^2 s_1^2 + w_2^2 s_2^2 + 2w_1 w_2 (r_{12} s_1 s_2) \quad (\text{II.B.14})$$

where s_p^2 is the variance of the returns to the portfolio,

s_i^2 is the variance of the returns to asset i ,

r_{12} is the correlation between the returns to assets 1 and 2,

w_i is the proportion invested in asset i .

The benefits of diversification are derived from adding assets to the portfolio that have low or even negative correlation with other assets in the portfolio, thus reducing the weighted sum of the correlations, and therefore the total risk of the portfolio. To demonstrate the risk-reducing effects of diversification, consider a portfolio consisting of 50% invested in the MSCI World Equity Index and 50% invested in the MSCI Sovereign Bond Index. From our descriptive statistics we know that the equity index has a variance of 74.98 and a standard deviation of 8.66. The bond index has a variance of 13.73 and a standard deviation of 3.71. In addition, the correlation coefficient is -0.163 .

If the asset returns had been perfectly correlated, then the portfolio volatility would have been as follows:

$$\begin{aligned} s_p &= \sqrt{(0.5^2 \times 74.98) + (0.5^2 \times 13.73) + 2(0.5 \times 0.5 \times 1 \times 8.66 \times 3.7)} \\ &= \sqrt{((0.5 \times 8.66) + (0.5 \times 3.7))^2} = (0.5 \times 8.66) + (0.5 \times 3.7) = 6.18\%. \end{aligned}$$

This shows that in this special case the portfolio risk is simply the weighted average standard deviation of the individual asset returns.

However, in other cases there is no such factorisation and hence no such simplification. When we apply the empirical correlation coefficient of -0.163 , the portfolio volatility becomes

$$s_p = \sqrt{(0.5^2 \times 74.98) + (0.5^2 \times 13.73) + 2(0.5 \times 0.5 \times (-0.163) \times 8.66 \times 3.7)} = 4.42\% .$$

Thus, because asset returns are, generally, not perfectly correlated, the standard deviation of a portfolio will be less than the weighted average of the standard deviations of the individual securities. Moreover, for given asset standard deviations, the standard deviation of a portfolio falls as the degree of correlation between pairs of assets falls.

Furthermore, repeating the computation with weights of 20% and 80% respectively gives a portfolio volatility of 3.18% , which is less than the volatility of each of the constituent assets.

II.C Calculus

Keith Parramore and Terry Watsham¹

The aim of this Chapter is to provide an introduction to the meaning of calculus and to its techniques, and to impart a foundation of skills and knowledge. By the end of this Chapter you will:

- understand the concept of *differentiation* and be able apply the rules of differentiation to polynomial, exponential and logarithmic functions;
- be able to apply *Taylor approximations* in financial contexts;
- understand the concept of *integration* and be able to apply the rules of integration to polynomial, exponential and logarithmic functions;
- be able to identify *minimum and maximum* values of univariate and multivariate functions;
- be able to solve unconstrained and constrained *optimisation* problems.

In this chapter we are concerned with two types of calculus, differential calculus and integral calculus.

Differential calculus enables us to measure the rate of change in one variable in relation to changes in one or more other variables. One common application in finance is measuring the rate of change of the price of a bond as a result of a change in the yield on that bond. Another frequently met application is the derivation of the structure of a portfolio of risky assets so as to maximise the portfolio return for a given level of portfolio risk. This latter application is known as *optimisation*.

Integral calculus enables us to find areas under curved lines or surfaces. Applications in finance include finding the expected value of a European option at expiry, and finding the probability of a given range of outcomes by finding the area under a probability density function. There are many problems which are similar to these examples, for instance finding the expectation and a lower percentile of a (profit and) loss distribution to estimate the (market or) credit value-at-risk of a portfolio (see Chapters III.A.2 and III.B.5).

¹ University of Brighton, UK.

Modern technology provides the means for computers to solve problems involving rates of change and areas/volumes, either by numerical methods or, in computer algebra packages, by analytic (exact) methods. Nevertheless, to use these tools the practitioner needs a good grounding in the underlying principles and techniques of calculus. So we will start our analysis with the study of differential calculus.

II.C.1 Differential Calculus

II.C.1.1 Functions

Suppose that a variable y changes as x changes so that for any given input x , there is a unique output, y . Then we say that y is a *function* of x , and write $y = f(x)$. To illustrate this consider the following functions:

$$y = 2x; \quad y = 5+2x; \quad y = 2x^2$$

Note that each of these functions is defined for any numerical value of x . This will not always be the case, and usually the specification of the *domain* of the function (the set on which the function operates) is an important part of the definition of a function. Let us calculate and sketch graphs for each of the above functions for values of x from -5 to 5 .

$y = 2x$	$x =$	-5	-4	-3	-2	-1	0	1	2	3	4	5
	$y =$	-10	-8	-6	-4	-2	0	2	4	6	9	10
$y = 5+2x$	$x =$	-5	-4	-3	-2	-1	0	1	2	3	4	5
	$y =$	-5	-3	-1	1	3	5	7	9	11	13	15
$y = 2x^2$	$x =$	-5	-4	-3	-2	-1	0	1	2	3	4	5
	$y =$	50	32	18	8	2	0	2	8	18	32	50

Figure II.C.1

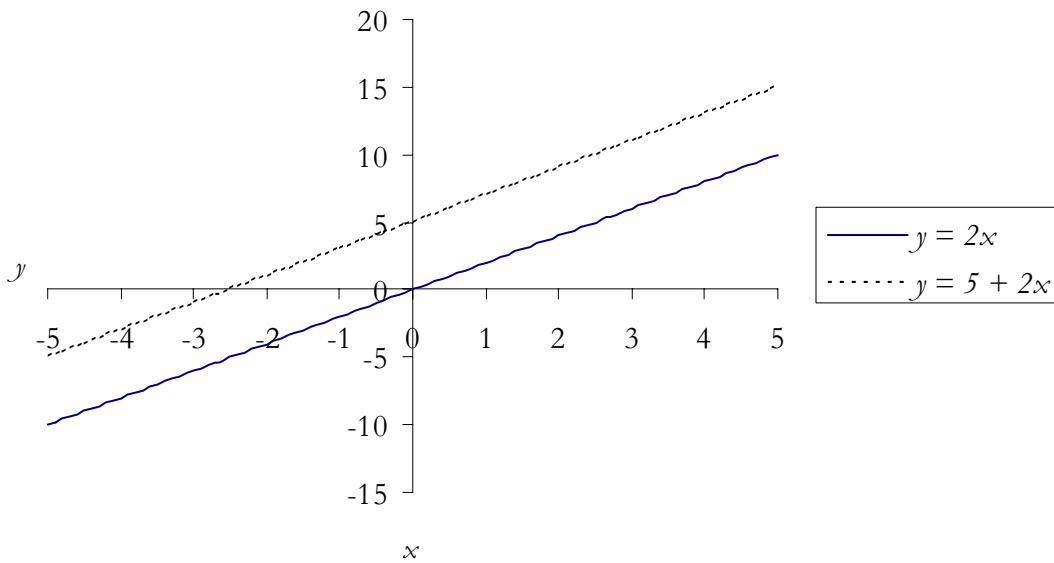
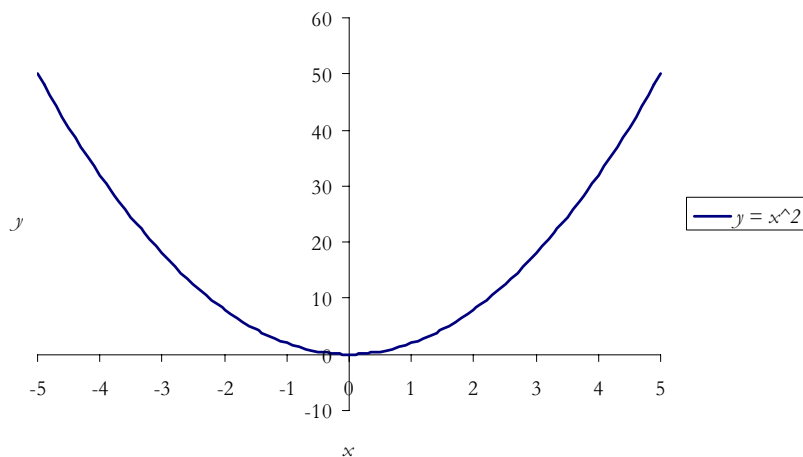


Figure II.C.2



II.C.1.2 The First Derivative

The first derivative of the function $y = f(x)$ (also stated as the first derivative of y with respect to x) shows how fast y is changing as a consequence of, and relative to, x changing. It shows the rate of change of y as x changes. There are two complementary notations for the first derivative,

$\frac{dy}{dx}$ and $f'(x)$. We will use both interchangeably.

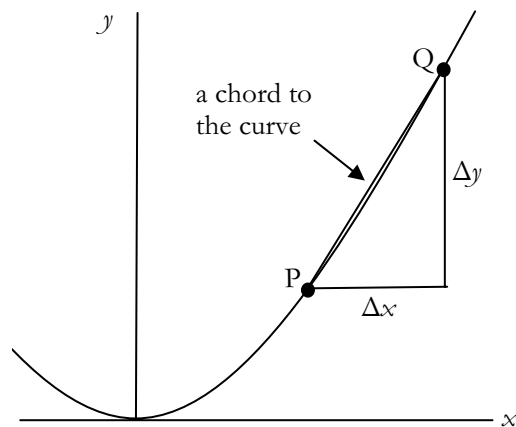
Let us first consider the function $y = 2x$ (Figure II.C.1). The slope of the line indicates the rate of change of y resulting from a change in x . The slope is given by the ratio of the vertical change

(i.e. the change in y , or Δy) divided by the horizontal change (i.e. the change in x , or Δx). Thus the slope is $\frac{\Delta y}{\Delta x}$. For the function $y = 2x$ this slope, the ratio of the vertical over the horizontal, is always the same. It has value 2. We get the same result with the function $y = 5 + 2x$. The change in y is always twice the change in x . For both functions $\frac{dy}{dx} = f'(x) = 2$. Note that the constant term, 5, in the second function does not influence the slope of the line, only its position. Thus the constant does not affect the rate of change and so does not influence the derivative of $f(x)$.

Now consider Figure II.C.2. Note that the curve gets steeper as x gets more positive. In other words the rate of change of y is not constant; it is increasing with x . In such situations, if x changes by a significant amount, say Δx , causing a change in y of Δy , then $f'(x)$ is only approximated by $\frac{\Delta y}{\Delta x}$.

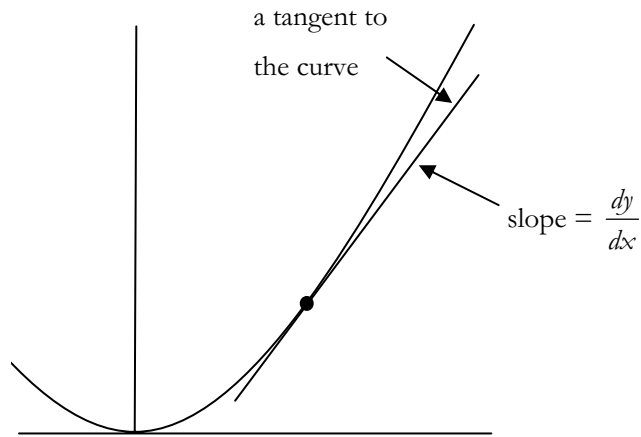
Consider Figure II.C.3. The hypotenuse of the right-angled triangle, labelled ‘a chord of the curve’, only gives an average rate of change between the points indicated.

Figure II.C.3



If Δx is made smaller and smaller, then the point Q approaches P, and the line through the points P and Q approaches the tangent at the point P. (Note that a *tangent* at any point on a curve is defined as a straight line that just touches the curve at that point). The gradient of the tangent is the rate of change of the function at point P. As Δx tends towards zero (Figure II.C.4), the difference quotient, $\frac{\Delta y}{\Delta x}$ tends to a function which gives that instantaneous rate of change.

Figure II.C.4



So how do we derive an expression for the instantaneous rate of change? Consider the function $y = 2x^2$. If x changes by a small amount, Δx , y will also change by a small amount, Δy . We get $y + \Delta y = 2(x + \Delta x)^2$. Expanding the right hand side gives $2x^2 + 4x\Delta x + 2(\Delta x)^2$. As $y = 2x^2$ we deduce that $\Delta y = 4x\Delta x + 2(\Delta x)^2$. Dividing both sides by Δx gives

$$\frac{\Delta y}{\Delta x} = \frac{4x\Delta x + 2(\Delta x)^2}{\Delta x} = 4x + 2\Delta x$$

This clearly tends towards $4x$ as Δx tends towards zero, so we conclude that

$$\frac{dy}{dx} = f'(x) = 4x.$$

The first derivative of the function $y = f(x) = 2x^2$ is thus $4x$. This gives the rate of change of y at any chosen point of x . It is the slope of the tangent at any chosen value of x .

II.C.1.3 Notation

The two notations for differentiation are due to Leibnitz and to Newton. The latter is also called functional notation.

Leibnitz notation If $y = 2x^2$ then $\frac{dy}{dx} = 4x$	Functional (Newton) notation If $f(x) = 2x^2$ then $f'(x) = 4x$
--	--

Leibnitz notation is useful when there is more than one independent (i.e. x) variable. It is then possible to refer separately to $\frac{dy}{dx_1}$ and $\frac{dy}{dx_2}$.

Functional notation is useful when there is a need to refer to a gradient at a specific point. Thus instead of ‘the value of $\frac{dy}{dx}$ at the point defined by $x = 2$ ’, reference can simply be made to $f'(2)$.

II.C.1.4 Simple Rules

Fortunately we do not have to go through all of the algebra every time we want to differentiate a function. There are some simple rules, as follows:

II.C.1.4.1 Differentiating Constants

$$\text{If } y = a \text{ (a constant), then } \frac{dy}{dx} = 0.$$

II.C.1.4.2 Differentiating a Linear Function

$$\text{If } y = bx, \text{ then } \frac{dy}{dx} = b.$$

II.C.1.4.3 The Gradient of a Straight Line

$$\text{If } y = a + bx, \text{ then } \frac{dy}{dx} = b.$$

II.C.1.4.4 The Derivative of a Power of x

$$\text{If } y = x^n, \text{ then } \frac{dy}{dx} = nx^{n-1}.$$

Example II.C.1:

(i) if $f(x) = 3x^5$ then $f'(x) = 3(5x^4) = 15x^4$

(ii) if $f(x) = x^{-4}$ then $f'(x) = -4(x^{-5}) = -4x^{-5}$, which is the same as:

if $f(x) = \frac{1}{x^4}$ then $f'(x) = -\frac{4}{x^5}$

(iii) if $f(x) = x^{1/3}$ then $f'(x) = \frac{1}{3}x^{-2/3} = \frac{1}{3x^{2/3}}$, which is the same as:

if $f(x) = \sqrt[3]{x}$ then $f'(x) = \frac{1}{3\sqrt[3]{x^2}}$

Example (ii) is particularly important in bond portfolio management because it is used to determine the interest rate sensitivity of bonds. This will be demonstrated later in this section.

II.C.1.4.5 Differentiating a scalar multiple of a function

If $y = k \times f(x)$, where k is a number, then $\frac{dy}{dx} = k f'(x)$.

II.C.1.4.6 Differentiating the Sum of Two Functions of x

Given the function $y = u + v$, where both u and v are functions of x , $\frac{dy}{dx} = \frac{du}{dx} + \frac{dv}{dx}$.

Example II.C.2:

(i) if $f(x) = 2x^4 + 3x^2$ then $f'(x) = 8x^3 + 6x$

(ii) if $f(x) = 3x^3 + 2x$ then $f'(x) = 9x^2 + 2$

II.C.1.4.7 Differentiating the Product of Two Functions of x

If $y = u \times v$, then $\frac{dy}{dx} = u \frac{dv}{dx} + v \frac{du}{dx}$.

So when y is a product of two functions of x , $\frac{dy}{dx}$ is found by multiplying each function by the derivative of the other function and adding the two products together.

Example II.C.3:

If $y = 4x^2(5x + 2)$, we can express the right-hand side as $u \times v$, where $u = 4x^2$ and $v = 5x + 2$. We

then have $\frac{du}{dx} = 8x$ and $\frac{dv}{dx} = 5$.

So $\frac{dy}{dx} = 4x^2 \times 5 + (5x + 2) \times 8x = 20x^2 + 40x^2 + 16x = 60x^2 + 16x$ or $4x(15x + 4)$.

Note that in this example the result can also be obtained by first multiplying out the bracket, i.e. $4x^2(5x + 2) = 20x^3 + 8x^2$, and then using earlier rules. This will not always be the case.

II.C.1.4.8 Differentiating the Quotient of Two Functions of x

If $y = \frac{u(x)}{v(x)}$, then $\frac{dy}{dx} = \frac{v \frac{du}{dx} - u \frac{dv}{dx}}{v^2}$.

Example II.C.4:

Assume again that $u = 4x^2$ and $v = 5x + 2$. As before, $\frac{du}{dx} = 8x$ and $\frac{dv}{dx} = 5$. So

$$= \frac{(5x + 2) \times 8x - 4x^2 \times 5}{(5x + 2)^2} = \frac{40x^2 + 16x - 20x^2}{(5x + 2)^2} = \frac{20x^2 + 16x}{(5x + 2)^2} \quad \text{or} \quad \frac{4x(5x + 4)}{(5x + 2)^2}$$

II.C.1.4.9 Differentiating a Function of a Function

$$\text{If } y = f(u), \text{ where } u = y(x), \text{ then } \frac{dy}{dx} = \frac{dy}{du} \times \frac{du}{dx}.$$

This is called the ‘chain rule’.

Example II.C.5:

If $y = (2x^3 + 3)^6$, then $y = u^6$, where $u = 2x^3 + 3$. So

$$\frac{dy}{du} = 6u^5 \text{ and } \frac{du}{dx} = 6x^2 \text{ and } \frac{dy}{dx} = 6u^5 \times 6x^2 = 36x^2(2x^3 + 3)^5.$$

II.C.1.4.10 Differentiating the Exponential Function

The exponential function e^x (note that it can also be written $\exp(x)$) is a particularly important function in calculus because, except for the zero function, this is the only function that differentiates to itself.

$$\text{If } y = e^x, \text{ then } \frac{dy}{dx} = e^x.$$

Example II.C.6:

- (i) if $f(x) = \exp(3x)$ then $f'(x) = 3\exp(3x)$ (using the chain rule)
- (ii) if $f(x) = \exp(x^2)$ then $f'(x) = 2x\exp(x^2)$ (again using the chain rule)
- (iii) if $f(x) = 5^x = (e^{\ln(5)})^x = e^{x \ln(5)}$ then $f'(x) = \ln(5)e^{x \ln(5)}$ (using the chain rule)
 $= \ln(5)(e^{\ln(5)})^x = \ln(5)5^x$

II.C.1.4.11 Differentiating the Natural Logarithmic Function

$$\text{If } y = \log_e(x) = \ln(x), \text{ then } \frac{dy}{dx} = \frac{1}{x}.$$

Example II.C.7:

If $f(x) = \ln(x^3 + 2)$ then $f'(x) = \frac{3x^2}{x^3 + 2}$ (using the chain rule).

Table II.C.1 summarizes the rules for carrying out differentiation.

Table II.C.1: Rules for differentiation

Rule	$y=f(x)$	$f'(x)$
1. constant	a	0
2. power rule	x^n e.g. x^2 x^3 x (i.e. x^1) 1 (i.e. x^0) $\frac{1}{x}$ (i.e. x^{-1}) $\frac{1}{x^2}$ (i.e. x^{-2}) \sqrt{x} (i.e. $x^{0.5}$)	nx^{n-1} $2x$ $3x^2$ 1 0 $-\frac{1}{x^2}$ (i.e. $-x^{-2}$) $-\frac{2}{x^3}$ (i.e. $-2x^{-3}$) $\frac{1}{2\sqrt{x}}$ (i.e. $0.5x^{-0.5}$)
3. constant multiple	$kf(x)$ e.g. $3x^5$	$kf'(x)$ $15x^4$
4. addition	$f(x) + g(x)$	$f'(x) + g'(x)$
5. product	$f(x) \times g(x)$	$f'(x)g(x) + f(x)g'(x)$
6. quotient	$\frac{f(x)}{g(x)}$	$\frac{g(x)f'(x) - f(x)g'(x)}{(g(x))^2}$
7. chain (composite functions)	$f(g(x))$	$g'(x) \times f'(g(x))$
8. exponential function	e^x (i.e. $\exp(x)$)	e^x
9. log function	$\ln(x)$ (i.e. $\log_e(x)$)	$\frac{1}{x}$

II.C.2 Case Study: Modified Duration of a Bond

The dirty price of a bond is the present value of all the future cash flows due under the bond. Assume that a series of cash flows, $CF(1)$, $CF(2)$, etc. are received at times 1, 2, etc. The present value of all of the cash flows (i.e. the dirty price) is PV , and the yield is y . This yield represents the yield per cash flow period (see Section I.B.2.4). For example, if the periods between cash flows were one year the yield would be an annual yield. However, if the periods were semi-annual the yields would be semi-annual yields. The yield is expressed as a decimal fraction, e.g. 8% is recorded as 0.08.

Note that in the English bond market, in which coupons are payable semi-annually, the annual yield is computed by doubling the semi-annual yield. Thus the annual yield is a *nominal* yield rather than an *effective* yield. For instance, a semi-annual yield of 5% would be quoted as an annual yield of 10%. But, by reinvesting, a semi-annual yield of 5% will deliver an effective annual yield of 10.25%, since $(1.05)^2 = 1.1025$.

The present value of the bond can be computed as²

$$PV = \frac{CF_1}{1+y} + \frac{CF_2}{(1+y)^2} + \dots + \frac{CF_n}{(1+y)^n} \quad (\text{II.C.1})$$

We are concerned here with the first derivative of the bond price with respect to yield. To obtain this we must differentiate each of the elements in the right-hand side of (II.C.1). Looking at the

first term, we can re-express it as $\frac{CF_1}{1+y} = CF_1 \times (1+y)^{-1}$. Using our rules for differentiation,

this differentiates to $-CF_1 \times (1+y)^{-2}$, which can be written as $-\frac{CF_1}{(1+y)^2}$. Similarly the first

derivative of the second term is $-\frac{2CF_2}{(1+y)^3}$ and in general the first derivative of the n th cash

flow is given by $-\frac{nCF_n}{(1+y)^{n+1}}$. Thus the first derivative of the bond price with respect to yield is:

$$\frac{dPV}{dy} = \sum_{i=1}^n \left(-\frac{iCF_i}{(1+y)^{i+1}} \right) = -\frac{CF_1}{(1+y)^2} - \frac{2CF_2}{(1+y)^3} - \dots - \frac{nCF_n}{(1+y)^{n+1}}.$$

This expression is identical to (I.B.2.8), except for the different notations used for the present value (P), cash flows (C or M) and the yield (r). This in turn can also be expressed as

² See also Section I.A.6.1

$$-\frac{1}{(1+y)} \left(\frac{CF_1}{1+y} + \frac{2CF_2}{(1+y)^2} + \dots + \frac{nCF_n}{(1+y)^n} \right). \quad (\text{II.C.2})$$

The proportional change in the bond price for a small change in the yield is given by:

$$\frac{1}{PV} \frac{dPV}{dy} = -\frac{1}{PV(1+y)} \left(\frac{CF_1}{1+y} + \frac{2CF_2}{(1+y)^2} + \dots + \frac{nCF_n}{(1+y)^n} \right). \quad (\text{II.C.3})$$

With a slight rearrangement of the right-hand side this becomes:

$$\frac{1}{PV} \frac{dPV}{dy} = -\frac{1}{1+y} \left(\frac{CF_1/(1+y)}{PV} + 2 \frac{CF_2/(1+y)^2}{PV} + \dots + n \frac{CF_n/(1+y)^n}{PV} \right). \quad (\text{II.C.4})$$

The expression in the bracket is a weighted sum of times. The times are the times at which a cash payment is received. For each time the weight is the proportion of the total present value represented by the present value of the corresponding cash payment. Thus the expression is an average time, the average (or expected) time at which cash is received.

The expression in the bracket is known as *Macaulay's duration*, after Macaulay (1938), and the whole expression is known as *modified duration*. Thus:

$$\frac{1}{1+y} \text{Macaulay's Duration} = \text{Modified Duration.}$$

More information about the duration of bonds is given in section I.B.2.6.

Example II.C.8:

Now let us consider a three-year bond paying annual coupons of 4 and trading on a yield-to-maturity of 5%. The dirty price of the bond is given as

$$PV = \frac{4}{1.05} + \frac{4}{1.05^2} + \frac{104}{1.05^3} = 97.2768.$$

The modified duration is

$$-\frac{1}{1.05} \left(\frac{4/1.05}{97.2768} + 2 \frac{4/1.05^2}{97.2768} + 3 \frac{4/1.05^3}{97.2768} \right) = -2.7470.$$

This is easy to evaluate in Excel:

	A	B	C	D	E	F	G	H	I	J
1		time		cashflow		present value		time*pv		Mod Dur
2		1		4		3.8095		3.8095		
3		2		4		3.6281		7.2562		
4		3		104		89.8391		269.5173		
5										
6						97.2768		280.5831		2.7470
7										

Note that if the cash flows arise at semi-annual intervals then the units of the modified duration will, correspondingly, be half-years. Thus it will then be necessary to divide by 2 to convert the modified duration into years.

Bond traders, portfolio managers and risk managers use modified duration through the following relationship:

$$\text{Percentage price change for a basis point change in yield} = \% \Delta P_B = \text{Modified Duration} \times 0.0001.$$

Thus in our example the bond is selling at 97.2768 with a modified duration of -2.7420 . We can compute the percentage change for a one basis point (0.01%) change in yield as $2.7420 \times 0.0001 = 0.00027420$. The actual price change explained by a one basis point change in yield = $0.00027420 \times 97.2768 = 0.02667$. This is known in the bond markets as the *PVBP* or the *PV01*.

This computation is important in the computation of value-at-risk for bond portfolios. In its simplest form, interest rate variability is represented by σ_y , the standard deviation of Δ_y , the daily change in yield. It then follows that σ_p , the standard deviation of the daily change in portfolio value, is given by:

$$\sigma_p = DP\sigma_y, \tag{II.C.5}$$

where D is the modified duration of the portfolio. Thus controlling D gives a methodology for controlling risk, and this is used in defining trading limits.

II.C.3 Higher-Order Derivatives

II.C.3.1 Second Derivatives

In Figure II.C.3 it is clear that the slope of the tangent changes as x changes. The rate at which the slope of that tangent is changing is the rate at which the rate of change is itself changing,

i.e. $\frac{d\left(\frac{dy}{dx}\right)}{dx}$. This is called the *second derivative* (or ‘second-order derivative’) of y with respect to x ,

and the notation is shortened to $\frac{d^2 y}{dx^2}$ or $f''(x)$. Here the second derivative is positive, meaning

that the slope is increasing. To find the second derivative, differentiate the first derivative. Thus

if $y = x^2$ then $\frac{dy}{dx} = 2x$ and $\frac{d^2 y}{dx^2} = 2$. The first derivative $\frac{dy}{dx}$ represents the rate of change of y

with respect to x , so $\frac{d^2 y}{dx^2}$ indicates whether that rate of change is increasing, constant or

decreasing.

II.C.3.2 Further Derivatives

We define higher-order derivatives in a similar way, though to avoid a confusion of dashes we

use, for instance, $f^{(3)}(x)$ instead of $f'''(x)$. Thus $f^{(3)}(x) = \frac{d}{dx} f''(x) = \frac{d}{dx} \left(\frac{d^2 y}{dx^2} \right) = \frac{d^3 y}{dx^3}$.

So if $f(x) = x^6$, say, then $f'(x) = 6x^5$, $f''(x) = 30x^4$, $f^{(3)}(x) = 120x^3$, $f^{(4)}(x) = 360x^2$, etc.

II.C.3.3 Taylor Approximations

We often know the present state of a process, and wish to know how that state might change as a consequence of a change in the underlying factors. For example, we may know the current price of a bond, but want to know how a bond price changes as the yield changes. Alternatively, we may want to know how the price of an option changes as the price of the underlying asset changes over time. When calculating value-at-risk it is more convenient to approximate the effect of a price change on a large portfolio than to do a complete portfolio revaluation. Approximations save computation time at the expense of accuracy. In some cases this trade-off is worth making.

In mathematical terms we want to know how $f(x)$ changes as x varies about its current value. In the examples just noted we can express the price of a bond as a function of the yield. Similarly, we can express the option price as a function of the asset price (and volatility, and time). The complexity of these functions makes approximation (rather than full valuation) attractive.

Below we give the 'first- and second-order Taylor approximations' of a function $f(x)$. (These approximations are only valid for *small* changes in the variable.) If we consider the function $f(x)$, we want to know how $f(x)$ changes as x changes by a small amount h .

The Linear (First-Order) Approximation

For this approximation we assume that we know, or can estimate, $f'(x)$, i.e. the gradient of $y = f(x)$ at our chosen value of x . The approximation is said to be of degree 1, and its graph is a straight line, the tangent to the curve. The intuition here is that we need to take account of the change in x , i.e. h , together with the rate of change of f with respect to x . Thus if $y = f(x)$ and x changes from its present value a by an amount h , then the consequential change in y is approximately $h \times f'(x)$.

This gives

$$f(x + h) \approx f(x) + f'(x) \times h. \quad (\text{II.C.6})$$

You can see that the approximation has the same value as $f(x)$ at the chosen value of x , and the same first derivative.

The Quadratic (Second-Order) Approximation

For the quadratic approximation we assume that we know, or can estimate, the second derivative $f''(x)$. The approximation is of degree 2, and therefore its graph is a parabola.

The approximation is chosen so that it has the same value of $f(x)$ at the chosen value of x , the same first derivative $f'(x)$ and the same second derivative $f''(x)$. It is defined to be

$$f(x+h) \approx f(x) + f'(x) \times h + \frac{1}{2} f''(x) \times h^2. \quad (\text{II.C.7})$$

Note: There is, in fact, a constant (zero-order) approximation, namely that

$$f(x+h) \approx f(x). \quad (\text{II.C.8})$$

There are also higher-order approximations, but we shall not be using them.

Summary of the Constant, Linear and Quadratic Taylor Series Approximations

The graphs in Figures II.C.5–II.C.7 show successive approximations applied to a bond yield curve. In each graph the bond yield curve is the solid line and the approximation is the dotted line. (Note that the relationship is $p = f(y)$, rather than $y = f(x)$.)

Figure II.C.5: Constant approximation $f(y + b) \approx f(y)$

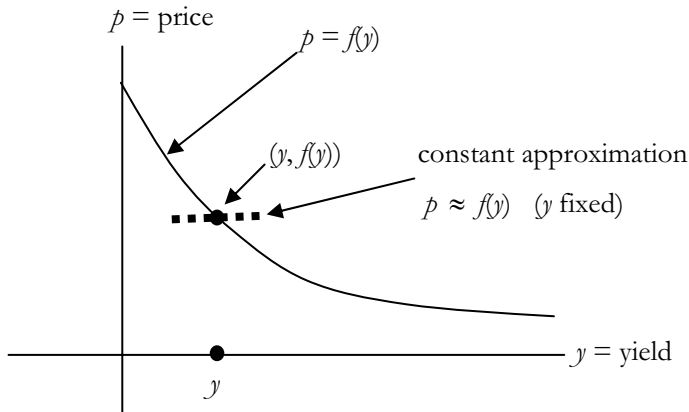


Figure II.C.6: Linear approximation $f(y + b) \approx f(y) + f'(y) \times b$

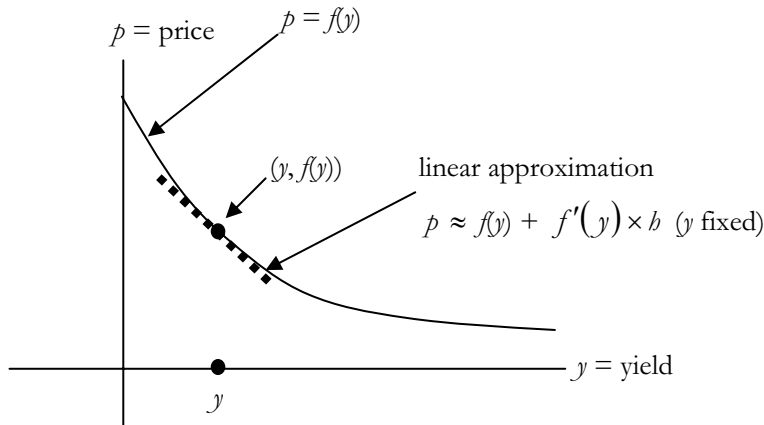
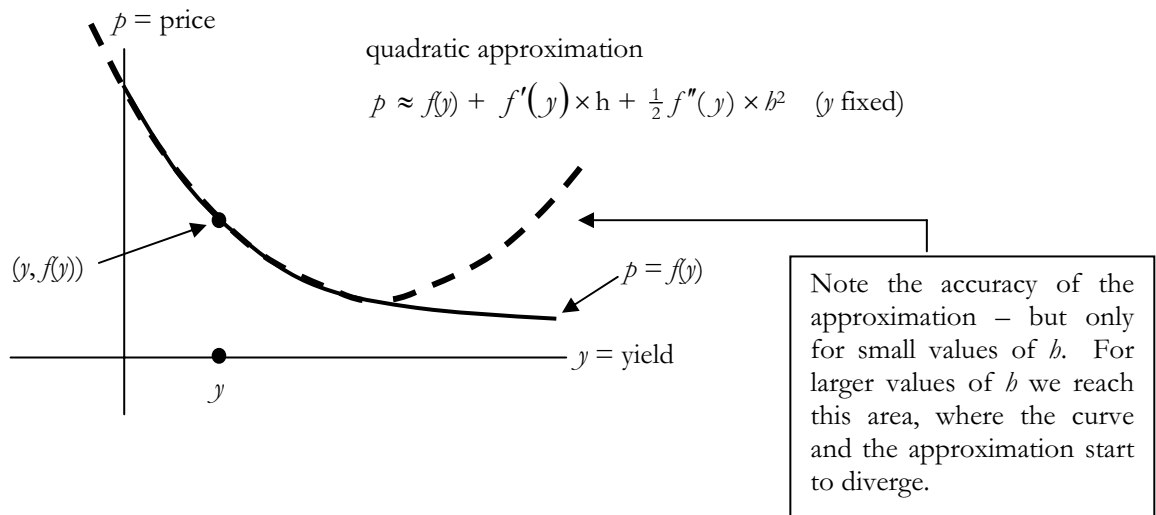


Figure II.C.7: Quadratic approximation $f(y + b) \approx f(y) + f'(y) \times b + \frac{1}{2} f''(y) \times b^2$



II.C.4 Financial Applications of Second Derivatives

II.C.4.1 Convexity

See also Chapter I.B.2.8. Recall that the standard deviation of the daily change in portfolio value, is given by $\sigma_p = DP\sigma_y$, where D is the modified duration of the portfolio (see Section II.C.2).

One application of the second derivative is to improve on our measure of bond price sensitivity given by modified duration through this equation. We can use our knowledge of the second derivative of a bond price with respect to the yield to maturity to calculate what is called bond *convexity*.

To illustrate this, recall that the derivative with respect to yield of a single cash flow is $-\frac{nCF_n}{(1+y)^{n+1}}$. Thus the second derivative is $\frac{n(n+1)CF_n}{(1+y)^{n+2}}$. Recalling the example of the three-year bond used in the duration example, the second derivative of the present value with respect to the yield is:

$$\frac{d^2PV}{dy^2} = \frac{2CF_1}{(1+y)^3} + \frac{6CF_2}{(1+y)^4} + \frac{12CF_3}{(1+y)^5}. \quad (\text{II.C.9})$$

Bond *convexity* is defined to be half the second derivative of the bond price with respect to yield, divided by the bond price, i.e. $(\partial^2 PV / \partial y^2) / 2PV$. Note that some practitioners refer to convexity as the second derivative divided by the present value, i.e. without the factor of $1/2$. The factor of $1/2$ is then introduced into the change-in-value computation.

Example II.C.9:

We will illustrate the use of the second derivative in bond risk management by calculating the convexity of the two-year bond we looked at earlier (Example II.C.8). Convexity is half the second derivative of the bond price with respect to yield, divided by the bond price. The second derivative of the bond price with respect to yield is:

$$\frac{2 \times 4}{1.05^3} + \frac{6 \times 4}{1.05^4} + \frac{12 \times 104}{1.05^5} = 6.9107 + 19.7447 + 977.84 = 1004.4962.$$

Given that the bond price is 97.2768, convexity is

$$\frac{1}{2} \times \frac{1004.4962}{97.2768} = 5.163$$

per cash flow period.

Modified duration is measured in years so convexity is measured in years squared. It is therefore necessary to divide the computed value by the square of the number of cash flows per year. In this example there is just one cash flow per year, so the convexity of this bond is 5.163 (years²). Note that if the cash flows arise at semi-annual intervals then it will then be necessary to divide by 4 to convert the convexity into (years²).

II.C.4.2 Convexity in Action

Now let us consider the three-year bond again. If the yield to maturity were to rise by 1% then the present value would fall from 97.2768 to 94.6540, a drop of 2.6228. The fall explained by modified duration would be $97.2768 \times 2.74703 \times 0.01 = 2.6722$; an overestimation of 0.0494. If we use the convexity measure as well as modified duration we will get a more accurate figure. The impact of convexity would be $97.2768 \times 5.1631 \times 0.01^2 = +0.0503$. So the combined effect attributable to duration and convexity of a 1% rise in yield would give a change of $-2.6722 + 0.0503 = -2.6219$, only 0.0009 in error relative to 2.6228. This is an example of how a second-order Taylor approximation can be better than a first-order approximation (see Section II.C.3.3.).

II.C.4.3 The Delta and Gamma of an Option

The value of an option is a function of the price of the underlying asset (and also of other variables). The first derivative of option value with respect to the price of the underlying is called the *delta* of the option; the second derivative is called the *gamma* of the option.

In terms of the Black–Scholes model (See I.A.8.7) the delta of a European call is $N(d_1)$, where N represents the cumulative (standard) normal density function and

$$d_1 = \frac{\ln\left(\frac{S}{X} + \left(r + \frac{\sigma^2}{2}\right)(T - t)\right)}{\sigma\sqrt{T - t}}.$$

The Black-Scholes gamma is given by $N'(d_1)/(S\sigma\sqrt{T-t})$ where N represents the (standard) normal density function. We then have that, if $W(S)$ is the value of a European call regarded as a function of S , the price of the underlying, then for a small change in that price δS ,

$$W(S + \delta S) \approx W(S) + \text{delta} \times \delta S + \frac{1}{2} \times \text{gamma} \times \delta S^2. \quad (\text{II.C.10})$$

This assumes that the volatility and the risk-free rate of interest are constant.

II.C.5 Differentiating a Function of More than One Variable

Under this heading we will consider two processes. The first differentiates a function consisting of several variables with respect to one of the variables, whilst holding all the others constant. The second differentiates a multivariate function assuming that all the variables change. The first is known as *partial differentiation* and the second is known as *total differentiation*.

II.C.5.1 Partial Differentiation

Consider the function $z = f(x, y)$, i.e. z is a function of two variables, x and y . Such a function can be differentiated with respect to one of the variables, while the other(s) are assumed held constant. This is known as *partial differentiation*.

Example II.C.10:

If $f(x, y) = x^2 + 6xy + 2y^3$ then the partial derivative of f with respect to x is $\frac{\partial f}{\partial x} = 2x + 6y$.

Similarly, the partial derivative of f with respect to y is $\frac{\partial f}{\partial y} = 6x + 6y^2$.

Note the notation (a type of delta rather than d) that is used to indicate a partial derivative, i.e. $\frac{\partial f}{\partial x}$ rather than $\frac{df}{dx}$.

An equation that contains partial differentials is known as a *partial differential equation* (PDE). These are particularly important for the valuation of derivative instruments. Specifically, the Black–Scholes PDE relates the value W of any derivative security to the price of the underlying, S , its volatility, σ , and the risk-free rate of return, r . It is:

$$\frac{\partial V}{\partial t} + rS \frac{\partial V}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} = rV. \quad (\text{II.C.11})$$

To understand the meaning of the terms of the Black–Scholes equation, first consider the following example. Suppose w is a function of three variables: $w = x^4 + xz^3 + xy - 4y^3$. Then

$$\frac{\partial w}{\partial x} = 4x^3 + z^3 + y, \quad \frac{\partial w}{\partial y} = x - 12y^2 \quad \text{and} \quad \frac{\partial w}{\partial z} = 3xz^2.$$

These results indicate that:

- As x is increased by a small amount Δx , w will increase by $(4x^3 + z^3 + y)\Delta x$.
- As y is increased by a small amount Δy , w will increase by $(x - 12y^2)\Delta y$.
- As z is increased by a small amount Δz , w will increase by $3xz^2\Delta z$.

Having established $\frac{\partial w}{\partial x}$, $\frac{\partial w}{\partial y}$ and $\frac{\partial w}{\partial z}$, it is natural to consider how *these* change with respect to

the underlying variables. For instance, the rate of change of $\frac{\partial w}{\partial x}$ with respect to x is $\frac{\partial\left(\frac{\partial w}{\partial x}\right)}{\partial x}$, which is shortened to $\frac{\partial^2 w}{\partial x^2}$. It reads ‘the second partial derivative of w with respect to x (twice)’.

The rate of change of $\frac{\partial w}{\partial x}$ with respect to y is $\frac{\partial\left(\frac{\partial w}{\partial x}\right)}{\partial y}$, which is shortened to $\frac{\partial^2 w}{\partial y\partial x}$. It reads ‘the second partial derivative of w with respect to x and y ’. Note that this turns out to be the same as $\frac{\partial^2 w}{\partial x\partial y}$, the second partial derivative of w with respect to y and x .

Thus for a function of three variables there are nine second partial derivatives. These are usually laid out as a symmetric matrix, known as the *Hessian* matrix (see Section II.D.3). In the above example the Hessian matrix is

$$\begin{bmatrix} \frac{\partial^2 w}{\partial x^2} & \frac{\partial^2 w}{\partial x\partial y} & \frac{\partial^2 w}{\partial x\partial z} \\ \frac{\partial^2 w}{\partial y\partial x} & \frac{\partial^2 w}{\partial y^2} & \frac{\partial^2 w}{\partial y\partial z} \\ \frac{\partial^2 w}{\partial z\partial x} & \frac{\partial^2 w}{\partial z\partial y} & \frac{\partial^2 w}{\partial z^2} \end{bmatrix}. \tag{II.C.12}$$

II.C.5.2 Total Differentiation

Total differentiation explains how the dependent variable will change when all the independent variables change. Consider again $w = f(x, y, z)$. For small changes in the independent variables

$\Delta w = \frac{\partial w}{\partial x} \Delta x$ and $\Delta w = \frac{\partial w}{\partial y} \Delta y$, where Δ represents a small change. Thus simultaneous small

changes in each of the variables gives

$$\Delta w = \frac{\partial w}{\partial x} \Delta x + \frac{\partial w}{\partial y} \Delta y + \frac{\partial w}{\partial z} \Delta z. \tag{II.C.13}$$

Example II.C.11:

If $w = x^4 + xz^3 + xy - 4y^3$, so that $\frac{\partial w}{\partial x} = 4x^3 + z^3 + y$, $\frac{\partial w}{\partial y} = x - 12y^2$ and $\frac{\partial w}{\partial z} = 3xz^2$, then the

total differential is $\Delta w = (4x^3 + z^3 + y)\Delta x + (x - 12y^2)\Delta y + 3xz^2\Delta z$.

Consider, for instance, $w(2, 1, 3)$ and $w(2.01, 1.05, 2.98)$.

$$w(2, 1, 3) = 16 + 54 + 2 - 4 = 68.$$

$$\Delta w = (32 + 27 + 1) \times 0.01 + (2 - 12) \times 0.05 + 54 \times (-0.02) = -0.98..$$

This would give 67.02 (= 68 - 0.98) as an approximation to $w(2.01, 1.05, 2.9)$. In fact the correct value of $w(2.1, 1.05, 2.9)$ is 66.9942 (to 4 decimal places).

Example II.C.12:

We have seen in Section II.C.4.3 that Taylor series approximations can be used to measure the change in value of an option when the underlying price changes. However, the value of an option also depends on the volatility of the underlying and on the risk-free rate of return. Considering just the first of these, the rate of change of option value with respect to the volatility of the underlying is known as vega, and is denoted by V .

The following approximation is often used (c.f. equation (II.D.7)):

$$\partial W \approx \Delta \partial S + V \partial \sigma + \frac{1}{2} \Gamma \partial S^2, \quad (\text{II.C.14})$$

where ∂W is the change in value of the option following a small change ∂S in the value of the underlying and a small change $\partial \sigma$ in its volatility. Δ is the portfolio's delta, V its vega and Γ its gamma.

This represents an example both of a Taylor expansion and of the total derivative in action. For instance, if we have an underlying with a volatility of 0.25, trading at 150 in an environment in which the risk-free rate of interest is 0.06, then a European call at a strike of 150 and with a life of 3 months has a value of 8.58953, a delta of 0.57240, a gamma of 0.02093 and a vega of 29.42652. (See Section II.D.2.3)

Thus if we take a portfolio consisting of just one of these calls, and consider what happens if the price of the underlying increases by 1 to 151 whilst its volatility increases by 0.02 to 0.27, then we can see that the call value changes by approximately

$$(0.57240 \times 1) + (29.42652 \times 0.02) + (0.5 \times 0.02093 \times 1) \approx 0.5724 + 0.5885 + 0.0105 \approx 1.171.$$

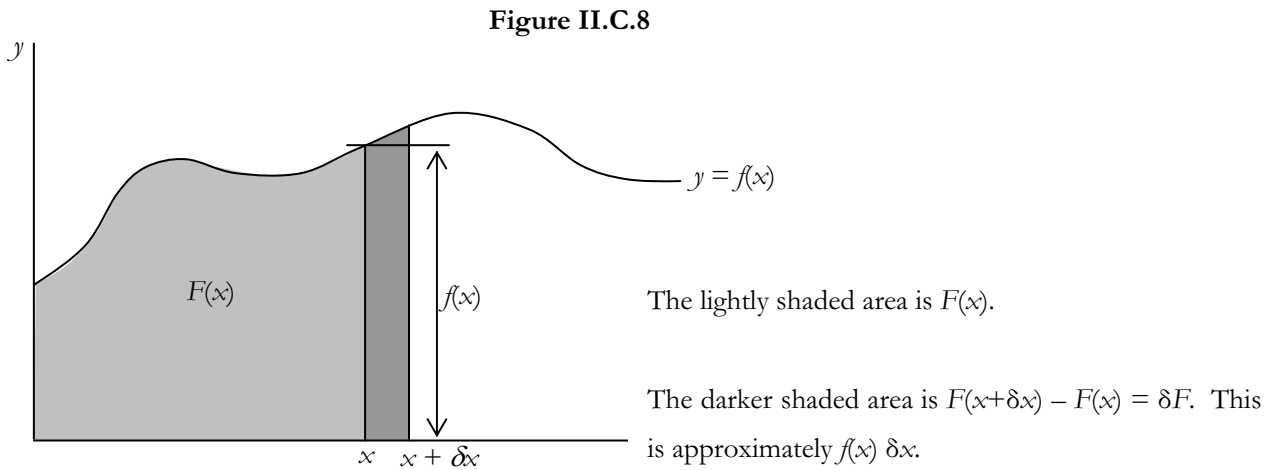
The corresponding Black–Scholes value is in fact 9.759, an increase of 1.169 on the original value of 8.590.

II.C.6 Integral Calculus

II.C.6.1 Indefinite and Definite Integrals

There are two aspects of integration for a function $f(x)$. The one is an intellectual exercise – finding a function which differentiates to the given function – which is called *indefinite integration*. It is also sometime called ‘anti-differentiation’. The second aspect is finding areas under the curve given by the function $f(x)$ using a *definite* integral between two values of x .

Figure II.C.8 illustrates the ‘fundamental theorem of analysis’ which, loosely speaking, says that definite and indefinite integration are two sides of the same coin. That is, if we can find a function $F(x)$ which differentiates to $f(x)$ we will be able to use it to find the area between $f(x)$ and the x -axis.



$$\text{So } \frac{\delta F}{\delta x} \cong f(x), \text{ and in the limit } \frac{dF}{dx} = f(x).$$

Note: For regions where $f(x)$ is negative the area between $f(x)$ and the x -axis is negative.

The process of producing a function which differentiates to f is called *integration*, and the notation is $\int f(x)dx$. This is called an *indefinite integral*.

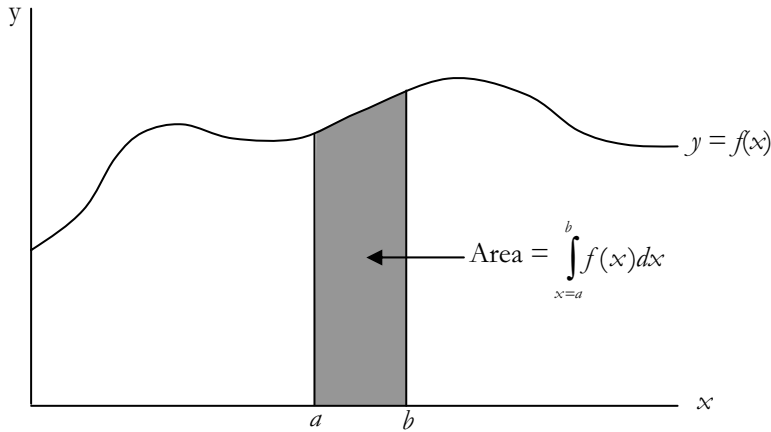
Figure II.C.9 illustrates the area under a curve described by a function $f(x)$. The area is $\int_{x=a}^b f(x)dx$, where a and b are the values of x at the left and right ends of the area respectively. This is called a *definite integral*.

The *fundamental theorem of analysis* states that, if we can find $F(x)$ so that $\frac{dF(x)}{dx} = f(x)$, then

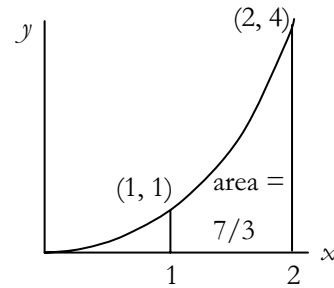
$$\int_{x=a}^b f(x) dx = F(b) - F(a).$$

This is written as $[F(x)]_a^b$.

Figure II.C.9



For instance, if $f(x) = x^2$ then $\int_{x=1}^{x=2} x^2 dx = \left[\frac{x^3}{3} \right]_1^2 = \frac{8}{3} - \frac{1}{3} = \frac{7}{3}$.



Differentiation is relatively easy. Integration is relatively hard. Experienced integrators do not remember integrals, they construct them from their knowledge of derivatives (see Section II.C.6.3). However, Table II.C.2 is included for your convenience.

Table II.C.2: A table of integrals

Function	indefinite integral
x^n	$\frac{x^{n+1}}{n+1}$ (except when $n = -1$)
x^2	$\frac{x^3}{3}$
x^3	$\frac{x^4}{4}$
x (i.e. x^1)	$\frac{x^2}{2}$
1 (i.e. x^0)	x
$\frac{1}{x}$ (i.e. x^{-1})	$\ln(x)$ (follows from $\frac{de^x}{dx} = e^x$)
$\frac{1}{x^2}$ (i.e. x^{-2})	$-\frac{1}{x}$ (i.e. $-x^{-1}$)
\sqrt{x} (i.e. $x^{0.5}$)	$\frac{2}{3}\sqrt[3]{x^3}$ (i.e. $\frac{x^{1.5}}{1.5}$)
e^x (i.e. $\exp(x)$)	e^x
$\ln(x)$ (i.e. $\log_e(x)$)	Advanced!

II.C.6.2 Rules for Integration

Corresponding to the rules for differentiation, there are rules for integration, which are given in Table II.C.3.

Table II.C.3: Rules for integration

Rule	$f(x)$	$\int f(x)dx$
1. constant multiple e.g.	$kf(x)$ $3x^5$	$k \int f(x)dx$ $3 \times \frac{x^6}{6} = \frac{x^6}{2}$
2. addition	$f(x) + g(x)$	$\int f(x)dx + \int g(x)dx$
3. parts	$f(x) \times g(x)$	$F(x)g(x) - \int F(x)g'(x)dx$
4. substitution	$f(g(x))$	replace $g(x)$ by u – but dx will also need replacing, using $du = g'(x)dx$, which can get messy!
4.5. substitution (special case)	$g'(x)f(g(x))$	$F(g(x))$ – You only have to integrate the f . The rest follows as a consequence of the chain rule for differentiation

II.C.6.3 Guessing

Experienced integrators use guessing and differentiation. Rule 1 allows you to 'guess and scale'. So to integrate $3x^5$, guess x^6 . This differentiates to $6x^5$ – two times too large. So halve the guess – rule 1 says that will make it right.

To integrate, say, e^{3x} , guess e^{3x} . Now $\frac{d}{dx}e^{3x} = 3e^{3x}$ (chain rule). This is three times too big, so

$$\int e^{3x} dx = \frac{e^{3x}}{3} \text{ (+ an arbitrary constant of integration).}$$

II.C.7 Optimisation

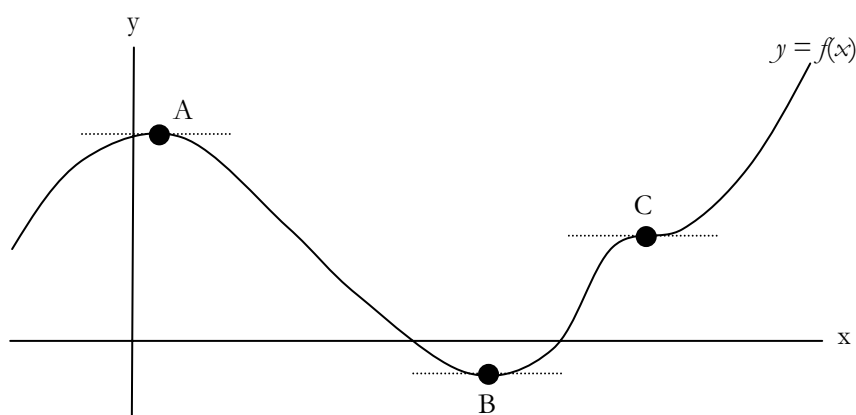
Optimisation is the process of finding the minimum or maximum value of a function. Recall that the first derivative identifies the rate of change of a function and that the second derivative indicates whether that rate of change is accelerating, slowing down or indeed stationary. If the maximum or minimum value has been reached then the function will, instantaneously, be neither increasing nor decreasing – it will be stationary. Consequently, we can use differential calculus, particularly the first derivative of a function, to find the ‘optimum’ points. The second derivative may then help us to see the nature of those points.

One application would be to find the structure of a portfolio that gives the minimum level of risk of all possible portfolios. We would call this problem ‘unconstrained’ optimisation. However, in finance we can rarely work without constraints. For instance, we would normally want to find the structure of the portfolio that gives the minimum risk, but subject to the constraint that the portfolio earns a minimum level of return. Given that there is generally a positive relationship between risk and return, structuring portfolios to have lower risk usually means having lower returns as well. Finding that structure which gives the lowest risk subject to a minimum level of return is a constrained optimisation problem.

In this section we will explain how to find the unconstrained minimum and maximum of a function of one variable and of a function of many variables. We will then explain the principles of finding the constrained optimum of a function of many variables. However, a practical application of constrained optimisation will have to wait until we have covered matrix algebra in Chapter II.D.

II.C.7.1 Finding the Minimum or Maximum of a Function of One Variable

Figure II.C.10



From Figure II.C.10 we can clearly see that point A is a local maximum. The tangent is horizontal, so the first derivative is zero. Point B is clearly a local minimum, and the first derivative is also zero. At point C the tangent is also horizontal, so it is a stationary point. But it is neither a maximum nor a minimum, it is a stationary point of inflection. The local maximum, the local minimum and the stationary point of inflection are all *stationary points*. These are points where $dy/dx = 0$.

If the local maximum, the local minimum and stationary points of inflection are all indicated by the first derivative being zero, how would we know which is which if we did not have the advantage of the diagram?

If at a stationary point the second derivative is negative, then we know that $dy/dx = 0$ and that dy/dx is falling. So as we move from left to right we must have a positive gradient followed by a zero gradient, followed by a negative gradient. This is characteristic of a local maximum, as at A.

A similar argument shows that if we have a stationary point at which the second derivative is positive, then we must have a gradient which, as we move from left to right, goes from negative to zero to positive. This is characteristic of a local minimum, as at B.

It would be neat if a second derivative of zero at a stationary point indicated a stationary point of inflection. Unfortunately this is not the case. For instance, at $x = 0$ the function $f(x) = x^4$ has derivative zero and second derivative zero, but it is a local minimum. All we can say if the second derivative is zero is that the second derivative test has failed. We then need to focus back on the first derivative to see whether or not it changes sign. If not, i.e. if it goes from positive to zero then back to positive as at C, or from negative to zero then back to negative, then we have a stationary point of inflection.

Example II.C.13:

Take for example the function $f(x) = \frac{54}{x} + x^2$ ($x \neq 0$)

The graph of this function $\left(y = \frac{54}{x} + x^2 \right)$ is shown

in Figure II.C.11.

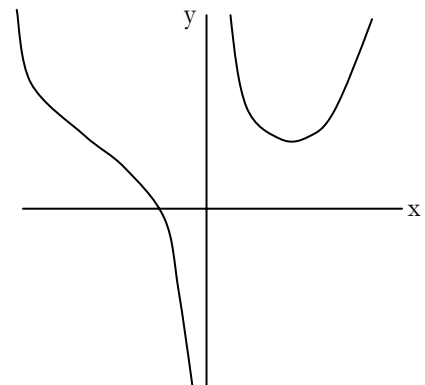


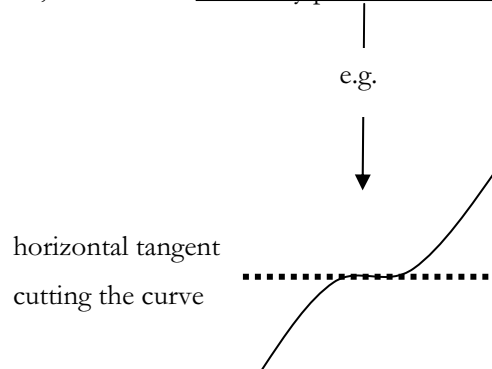
Figure II.C.11

The first stage in finding the maximum or minimum value of a function is to set the first derivative to zero. Differentiating gives $\frac{dy}{dx} = -\frac{54}{x^2} + 2x$. This is equal to zero when $x = 3$.

The graph shows that this is a minimum if we restrict ourselves to positive values of x , but if we take all values the function has no minimum. For small negative values of x the function value is *very* negative, i.e. negative and very large in magnitude. The minimum in the ‘domain’ of positive numbers is what is known as a ‘local’ minimum. In this particular example there is no global minimum.

In summary, the technique for finding optimum points is:

1. Find stationary points, i.e. points where $f'(x) = 0$.
2. Classify the stationary points into maxima, minima and stationary points of inflection.



A failsafe way to classify stationary points is by considering how the derivative behaves (the first derivative test), but the second derivative test can be useful:

First derivative test

x	just less than 3	3	just more than 3
$\frac{dy}{dx} = -\frac{54}{x^2} + 2x$	< 0	0	> 0
function shape	\	—	/

The shape shows that we have a local minimum.

Second derivative test

Differentiate again to find $\frac{d^2y}{dx^2}$ and evaluate it at the stationary point.

If the value is positive then the stationary point is a (local) minimum.

If the value is negative then the stationary point is a local maximum.

If the value is zero then the second derivative test fails.

Note: If the value of the second derivative is zero then it is *not* the case that the stationary point is necessarily a stationary point of inflection. For instance, consider $y = x^4$ at $x = 0$.

So $f'(x) = -\frac{54}{x^2} + 2x$ gives $f''(x) = \frac{108}{x^3} + 2$.

$f'(3) = 0$ and $f''(3) = 14$, so there is a local minimum at $x = 3$.

II.C.7.2 Maxima and Minima of Functions of More than One Variable

A stationary point of a function of more than one variable is a point where all partial derivatives are zero. A stationary point can be a local minimum or a local maximum or a saddle point. The latter is a point with a maximum in at least one direction, and a minimum in at least one direction. A saddle is one example, and a second is a mountain col, where the function is a function of two position variables (such as latitude and longitude) and the function value is height.

A local minimum may be a *strong* local minimum (the bottom of a bowl) or may be *weak* (a point on the horizontal floor of a trough). Similarly, a local maximum may be strong or weak.

To discover the type of a stationary point the Hessian matrix of second partial derivatives is examined (see Section II.C.5.1). Mirroring the situation with one variable, if at a stationary point the Hessian of second derivatives is positive definite (see Section II.D.3.2) then we have a (local) minimum; if it is negative definite then we have a (local) maximum; if it is neither then we do not know, it could be a local maximum, a local minimum, or a saddle point – we would need to examine the change in the gradient to see which.

The two-variable case is easy to visualize since the function value can be seen as corresponding to height, the location being determined by two variables such as latitude and longitude. Consider such a function $f(x, y)$. This has two partial first derivatives and four partial second derivatives. They are:

$$\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \text{ and } \frac{\partial^2 f}{\partial x^2}, \frac{\partial^2 f}{\partial y^2}, \frac{\partial^2 f}{\partial x \partial y}, \frac{\partial^2 f}{\partial y \partial x} \left(\text{but note that } \frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x} \right).$$

In the two-variable situation, there are three types of stationary point, local maxima, local minima and saddle points. The latter are like the high points of mountain passes – mountain peaks on each side and valleys in front and behind. The conditions for a strong local extremum are as follows.

$$\text{Local maxima: } \frac{\partial f}{\partial x} = \frac{\partial f}{\partial y} = 0; \quad \frac{\partial^2 f}{\partial x^2} < 0; \quad \frac{\partial^2 f}{\partial x^2} \frac{\partial^2 f}{\partial y^2} > \left(\frac{\partial^2 f}{\partial x \partial y} \right)^2. \quad (\text{II.C.15})$$

$$\text{Local minima: } \frac{\partial f}{\partial x} = \frac{\partial f}{\partial y} = 0; \quad \frac{\partial^2 f}{\partial x^2} > 0; \quad \frac{\partial^2 f}{\partial x^2} \frac{\partial^2 f}{\partial y^2} > \left(\frac{\partial^2 f}{\partial x \partial y} \right)^2. \quad (\text{II.C.16})$$

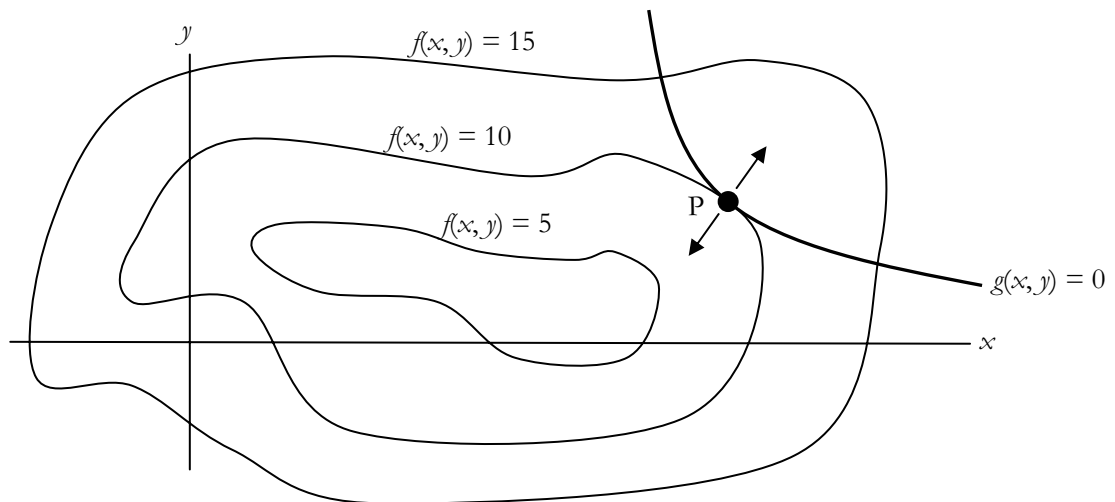
For the conditions for weak local extrema, which allow ridges as well as peaks, or valley bottoms as well as bottoms of bowls, change the $>$ and $<$ to \leq and \leq respectively.

II.C.7.3 Optimization Subject to Constraints: Lagrange Multipliers

There are many examples in business and finance where we wish to find the maximum or minimum of a function subject to a constraint. For example, in portfolio management we frequently wish to know the minimum risk that is achievable subject to a required expected return. In constrained optimisation we use what is known as a ‘Lagrange multiplier’.

Suppose we wish to maximize the objective function, $f(x, y)$, subject to the constraint $g(x, y) = 0$.³ We might, for instance be trying to maximise the return to a portfolio given an acceptable level of risk (see Section II.D.2).

Again, using the fact that we can visualize a function of two variables as a height, we can draw contours of f , just as on a map. We can then draw $g(x, y) = 0$ on the same graph. The result might look something like this:



We can see that the constrained minimum is at P, and furthermore that the contours of f and g are tangent at this point. This must be so since otherwise the line $g(x, y) = 0$ would be crossing

³ Note that the situation is more complex when the constraints are inequalities rather than equalities. That situation is dealt with by Kuhn–Tucker analysis; see also Section II.G.2.3.

the f -contour, which would mean that there would be a point ‘downhill’ from P. That is, there would be a point on the g -contour with a lower value of f . Thus the vector of partial derivatives for f must be in the same direction (positive or negative, as indicated by the arrows at P) as the vector of partial derivatives for g . In two dimensions this means that the two partial derivatives of f must be in the same proportion as the two partial derivatives of g . This can be expressed as

$$\frac{\partial f / \partial y}{\partial f / \partial x} = \frac{\partial g / \partial y}{\partial g / \partial x}.$$

To capture this in this simple situation and in more complex situations, where there may be more than one constraint and where inequalities may be involved, the *Lagrangian* is constructed as follows:

$$L(x, y, \lambda) = f(x, y) - \lambda g(x, y). \quad (\text{II.C.17})$$

In this case the Lagrangian is a function of three variables, x , y and λ , where λ is known as the *Lagrange multiplier* and is constant.

We now consider the partial derivatives of the Lagrangian:

$$\frac{\partial L}{\partial x} = \frac{\partial f}{\partial x} - \lambda \frac{\partial g}{\partial x}; \quad \frac{\partial L}{\partial y} = \frac{\partial f}{\partial y} - \lambda \frac{\partial g}{\partial y}; \quad \frac{\partial L}{\partial \lambda} = -g.$$

Setting the first two of these to zero gives the condition that the gradients are in the same direction, i.e. that

$$\frac{\partial f / \partial y}{\partial f / \partial x} = \frac{\partial g / \partial y}{\partial g / \partial x}.$$

Setting the third partial derivative to zero gives the constraint $g(x, y) = 0$. Hence to solve the constrained optimization problem we look for a stationary point of the Lagrangian, i.e. a point where the partial derivatives are all equal to zero.

Example II.C.14:

To illustrate this, suppose that we wish to maximize the expression $5x + 2x^2 + 4y$ subject to the constraint that $2x + y = 20$.

In this example $f(x, y) = 5x + 2x^2 + 4y$ and $g(x, y) = 2x + y - 20$. (Note the reorganisation that is needed so that the constraint is in the form $g(x, y) = 0$.) The Lagrangian is then

$$L(x, y, \lambda) = 5x + 2x^2 + 4y - \lambda(2x + y - 20).$$

Differentiating:

$$\frac{\partial L}{\partial x} = 5 + 4x - 2\lambda; \quad \frac{\partial L}{\partial y} = 4 - \lambda; \quad \frac{\partial L}{\partial \lambda} = 2x + y - 20.$$

Setting each to zero gives

$$5 + 4x - 2\lambda = 0 \quad 4 - \lambda = 0 \quad 2x + y - 20 = 0.$$

Solving gives $x = 0.75$, $y = 18.5$ and $\lambda = 4$. Thus the constrained maximum is $f(0.75, 18.5) = 78.875$.

Note that the value of λ is, in economic parlance, the ‘marginal value’ of relaxing the constraint. So if the constraint were to be changed to $2x + y = 20 + b$, the constrained maximum would change from 78.875 to $78.875 + 4b$, provided that b is small enough not to cause any ‘structural’ change to the solution.

II.C.7.4 Applications

The standard Markowitz (1952) portfolio optimisation problem is to determine how to spread an investment across a number of assets so as to achieve a required expected return with minimum risk. This is a constrained optimisation problem. The problem is solved in practice using commercial optimisation packages, and the background is covered in Section II.D.2, where it is solved using Excel’s optimiser, which is called ‘Solver’. Most modern portfolio management techniques are now based on advanced constrained optimisation problems, where constraints can range from simple ‘no short sales’ constraints to advanced constraints that limit turnover costs, re-balancing frequency and so forth. For the application of optimisation to regression, see Chapter II.F. Beyond asset management, optimisation is a core technique for calibration of option pricing models, for instance with the minimization of errors between model and market prices.

References

Markowitz, H (1952) Portfolio selection, *Journal of Finance*, 7(1), pp. 77–91.

II.D Linear Mathematics and Matrix Algebra

Keith Parramore and Terry Watsham¹

After studying this chapter you will understand vector and matrix algebra, including: addition and subtraction; matrix multiplication; the transpose and inverse of square matrices; special types of matrices and the laws of matrix algebra; the Cholesky decomposition of a matrix; and eigenvalues and eigenvectors. In addition, you will be able to apply this understanding to common financial applications such as manipulating covariance matrices, calculating the variance of the returns to a portfolio of assets, hedging a vanilla option position, and simulating correlated sets of returns.

In examining the returns and risks involved in holding portfolios of securities we find that both linear and nonlinear mathematics are involved. Surprisingly, techniques associated with linear mathematics, specifically matrix algebra, simplify the handling of the nonlinear aspects. We will see this in action when we deal with the variance of the returns to a portfolio. That variance is a quadratic function of the asset weightings, but we deal with it by using the covariance matrix (Section II.D.2). We will be using matrices to solve simultaneous equations. This will be needed in the portfolio work, and will also be useful elsewhere.

A function or a process is said to be linear if, for instance, twice the input produces twice the output, and if the output from the sum of two inputs is the sum of their individual outputs. Thus linearity encapsulates *proportionality*. As an example, consider the return to a portfolio of two assets. In Chapter II.B you saw that this is a linear function of the returns to the individual assets:

$$R_P = wR_1 + (1 - w)R_2,$$

where R_P represents the (discrete) return to the portfolio, R_1 and R_2 represent the returns to the assets, and w and $1 - w$ represent the proportions invested in each asset. On the other hand, the standard deviation of the returns is nonlinear since the standard deviation of the returns to the portfolio is the square root of a quadratic function of w and the standard deviations and correlation of the returns to the individual assets. In general:

$$F \text{ linear} \Leftrightarrow F(\lambda \mathbf{x} + \mu \mathbf{y}) = \lambda F(\mathbf{x}) + \mu F(\mathbf{y}),$$

¹ University of Brighton, UK.

where Greek letters represent numbers (i.e., scalars) and where bold letters represent vectors (see below). Whilst there are many functions and processes that are linear, there are many others that are not. For instance, the natural logarithm function is not a linear function:

$$\ln(1) = 0, \ln(e) = 1, \ln(1 + e) \approx 1.313262, \text{ so } \ln(1 + e) \neq \ln(1) + \ln(e).$$

II.D.1 Matrix Algebra

II.D.1.1 Matrices

A matrix is a rectangular array of numbers. Its shape (also known as its *order*) is given as the number of rows and the number of columns. This is written as $r \times c$. Thus

$$\begin{bmatrix} 2 & 5 & 1 \\ 0 & 6 & 4 \end{bmatrix} \text{ is a } 2 \times 3 \text{ matrix.}$$

The individual cells in a matrix are identified first by their position in the row, and then by their position in the column. Thus, the 2 at the left-hand end of the first row in the above matrix is at cell number 1,1, and the 4 at the right hand end of the bottom row is at cell number 2,3.

In their algebraic form double-digit subscripts are often used to identify an element within a matrix. For example, the element at the junction of the i th row and j th column of matrix \mathbf{Y} would be given as y_{ij} . The matrix itself might be written as $\mathbf{Y} = [y_{ij}]$.²

Matrices can be multiplied if their shapes satisfy particular conditions. They can be added and subtracted, but only if each matrix is of the *same order*. Matrices cannot be divided, but, as we will see later, multiplying matrix \mathbf{A} by the inverse of matrix \mathbf{B} is similar to dividing matrix \mathbf{A} by matrix \mathbf{B} . It is usual to give matrices letter names, and those letters are often shown in bold type.

II.D.1.2 Vectors and Transposes

Vectors are columns of data, that is, matrices with only one column. Thus a *vector* (or column vector) is a matrix of order $n \times 1$. A *row vector* is a matrix of order $1 \times n$. If the vector has a letter name then the letter is often shown in lower-case bold type, for example,

$$\mathbf{x} = \begin{bmatrix} 1 \\ 5 \\ 3 \end{bmatrix}.$$

² Note that most users would regard a 1×1 matrix as a number, or scalar. But we shall see later that Excel draws a distinction between a scalar and a 1×1 matrix.

A matrix produced by swapping the rows and columns of a given matrix is called the *transpose* of that matrix. The transpose of a (column) vector is a row vector. A transpose is indicated by the T superscript (or a ‘prime’ symbol $'$ – in this chapter we use the former). For example,

$$\begin{bmatrix} 3 & 12 & 9 & 29 \\ 2 & 18 & 4 & 61 \end{bmatrix}^{\text{T}} = \begin{bmatrix} 3 & 2 \\ 12 & 18 \\ 9 & 4 \\ 29 & 61 \end{bmatrix} \text{ and } \begin{bmatrix} 2 \\ 8 \\ 1 \end{bmatrix}^{\text{T}} = [2 \ 8 \ 1].$$

II.D.1.3 Manipulation of Matrices

Before matrix algebra (i.e., the ways in which matrices can be combined) can be put to use, we must study how matrices can be used to model financial problems. We will explain that later in this chapter. Here we will continue by explaining the addition and subtraction of matrices.

Matrix Addition and Subtraction

Matrices can only be added or subtracted if they are of the same order. Matrices are added together by adding each element in one matrix to the corresponding element in the other matrix. Subtraction of matrices is achieved by subtracting each element in the second matrix from the corresponding element in the first. Thus

$$\mathbf{X} + \mathbf{Y} = [x_{ij} + y_{ij}] \text{ and } \mathbf{X} - \mathbf{Y} = [x_{ij} - y_{ij}]. \quad (\text{II.D.1})$$

Example II.D.1:

$$\begin{bmatrix} 3 & 4 & 2 \\ 0 & 6 & 5 \end{bmatrix} + \begin{bmatrix} 2 & -2 & 3 \\ 5 & 5 & 1 \end{bmatrix} = \begin{bmatrix} (3+2) & (4-2) & (2+3) \\ (0+5) & (6+5) & (5+1) \end{bmatrix} = \begin{bmatrix} 5 & 2 & 5 \\ 5 & 11 & 6 \end{bmatrix}$$

and

$$\begin{bmatrix} 3 & 4 & 2 \\ 0 & 6 & 5 \end{bmatrix} - \begin{bmatrix} 2 & -2 & 3 \\ 5 & 5 & 1 \end{bmatrix} = \begin{bmatrix} (3-2) & (4-(-2)) & (2-3) \\ (0-5) & (6-5) & (5-1) \end{bmatrix} = \begin{bmatrix} 1 & 6 & -1 \\ -5 & 1 & 4 \end{bmatrix}.$$

Figure II.D.1 shows how to carry out matrix addition in Excel.

Scalar Multiplication

We have to distinguish between the multiplication of a matrix by a scalar and multiplication by another matrix. The former is known as *scalar multiplication*, and involves multiplying every element

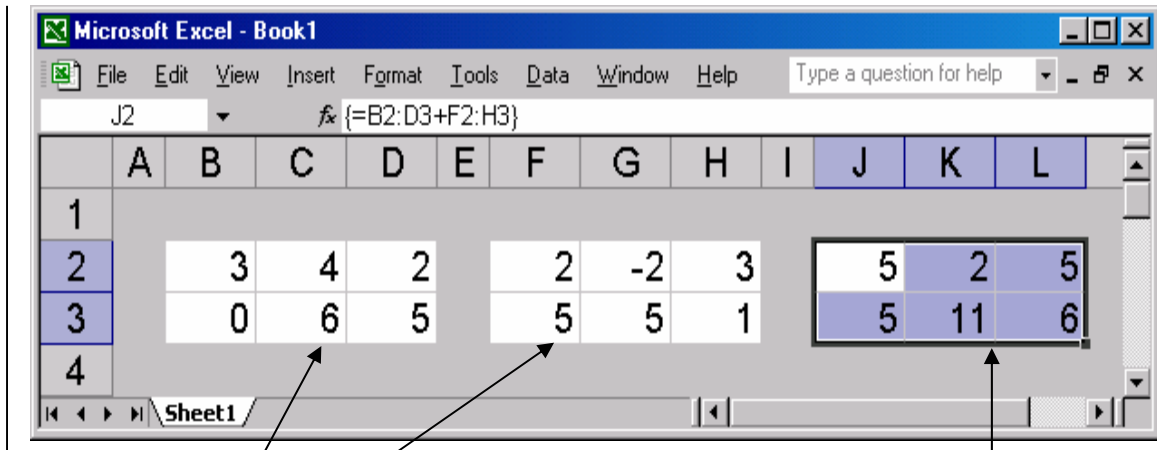
of the matrix by that number. Thus, for instance if matrix \mathbf{X} is $\begin{bmatrix} 8 & 6 \\ 3 & 2 \end{bmatrix}$, then

$$2\mathbf{X} = 2 \begin{bmatrix} 8 & 6 \\ 3 & 2 \end{bmatrix} = \begin{bmatrix} 16 & 12 \\ 6 & 4 \end{bmatrix}.$$

Sumproduct

There are many instances when we have two paired sets of numbers which we need to multiply together in their pairs and then sum the results – for example, the quantities of a number of items together with their prices. Again, Excel has a useful facility to handle this (Figure II.D.2).

Figure II.D.1: [Matrix addition in Excel](#)



These matrices were entered as numbers in their individual cells.

To perform the matrix addition these six cells were first selected.

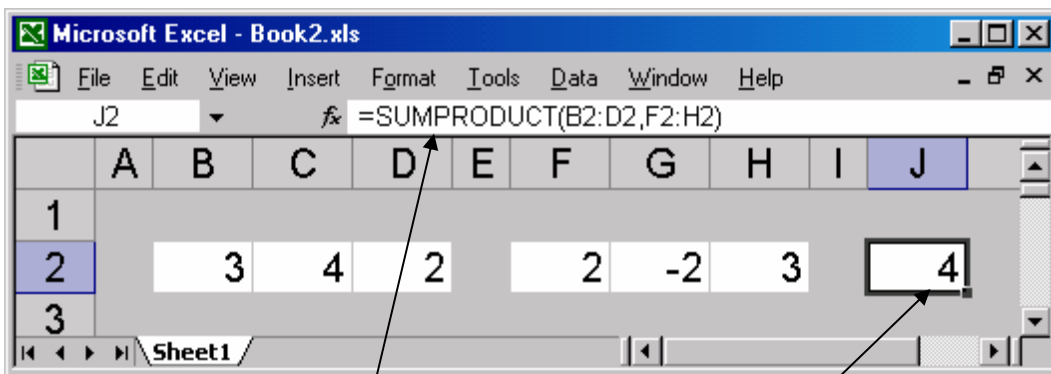
The following formula was then typed: =B2:D3+F2:H3

Finally the formula was entered by holding down the **Ctrl** and the **Shift** keys whilst hitting the **Enter** key.

All six cells are then identified as a matrix, this being indicated in the formula bar by: {= B2:D3+F2:H3}

(Note the { ... } brackets.)

Figure II.D.2: [Sumproduct in Excel](#)



The formula bar shows that the formula behind this cell is:
 =SUMPRODUCT(B2:D2, F2:H2), giving $3 \times 2 + 4 \times (-2) + 2 \times 3 = 4$

II.D.1.4 Matrix Multiplication

Matrices can also be multiplied together, provided that the number of columns of the first matrix is equal to the number of rows of the second, that is, provided that their orders are of the form $a \times b$ and $b \times c$.

Example II.D.2:

If $\mathbf{X} = \begin{bmatrix} 6 & 2 & 1 \\ 8 & 9 & 4 \end{bmatrix}$ (order 2×3) and $\mathbf{Y} = \begin{bmatrix} 2 & 8 & 4 & 0 \\ 3 & 4 & 2 & 3 \\ 1 & 6 & 3 & 0 \end{bmatrix}$ (order 3×4) then these two can be

multiplied, as the number of columns in \mathbf{X} (namely 3) is equal to the number of rows in \mathbf{Y} (also 3). Therefore, $\mathbf{Z} = \mathbf{XY}$ can be computed, and is of order 2×4 (the number of rows of \mathbf{X} \times the number of columns of \mathbf{Y}).

In general, multiplying a matrix of order $a \times b$ with a matrix of order $b \times c$ results in a matrix of order $a \times c$, as will be seen below. Multiplying together our matrices \mathbf{X} and \mathbf{Y} from above is effectively applying a number of sumproduct operations (Section II.D.1.3), as illustrated below.

$$\begin{array}{ccc} \begin{bmatrix} 6 & 2 & 1 \\ 8 & 9 & 4 \end{bmatrix} & \begin{bmatrix} 2 & 8 & 4 & 0 \\ 3 & 4 & 2 & 3 \\ 1 & 6 & 3 & 0 \end{bmatrix} & = & \begin{bmatrix} z_{11} & z_{12} & z_{13} & z_{14} \\ z_{21} & z_{22} & z_{23} & z_{24} \end{bmatrix} \\ \mathbf{X} & \times & \mathbf{Y} & = & \mathbf{Z} \end{array}$$

Thus to calculate the first element (z_{11}) of the matrix \mathbf{Z} we obtain the sumproduct of the first row of \mathbf{X} with the first column of \mathbf{Y} , $(6 \times 2) + (2 \times 3) + (1 \times 1) = 19$. Other elements in the top row of \mathbf{Z} (z_{12} , z_{13} and z_{14}) are computed similarly by using the first row of \mathbf{X} and the second, third and fourth columns of \mathbf{Y} , respectively: The results are

$$\begin{aligned} z_{12} &= (6 \times 8) + (2 \times 4) + (1 \times 6) = 62, \\ z_{13} &= (6 \times 4) + (2 \times 2) + (1 \times 3) = 31, \\ z_{14} &= (6 \times 0) + (2 \times 3) + (1 \times 0) = 6. \end{aligned}$$

The subsequent row(s) of the \mathbf{Z} matrix are computed from subsequent row(s) of \mathbf{X} and all the appropriate columns of \mathbf{Y} . This gives

$$\mathbf{Z} = \begin{bmatrix} 19 & 62 & 31 & 6 \\ 47 & 124 & 62 & 27 \end{bmatrix}.$$

Note that *matrix multiplication is associative*, that is,

$$\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}.$$

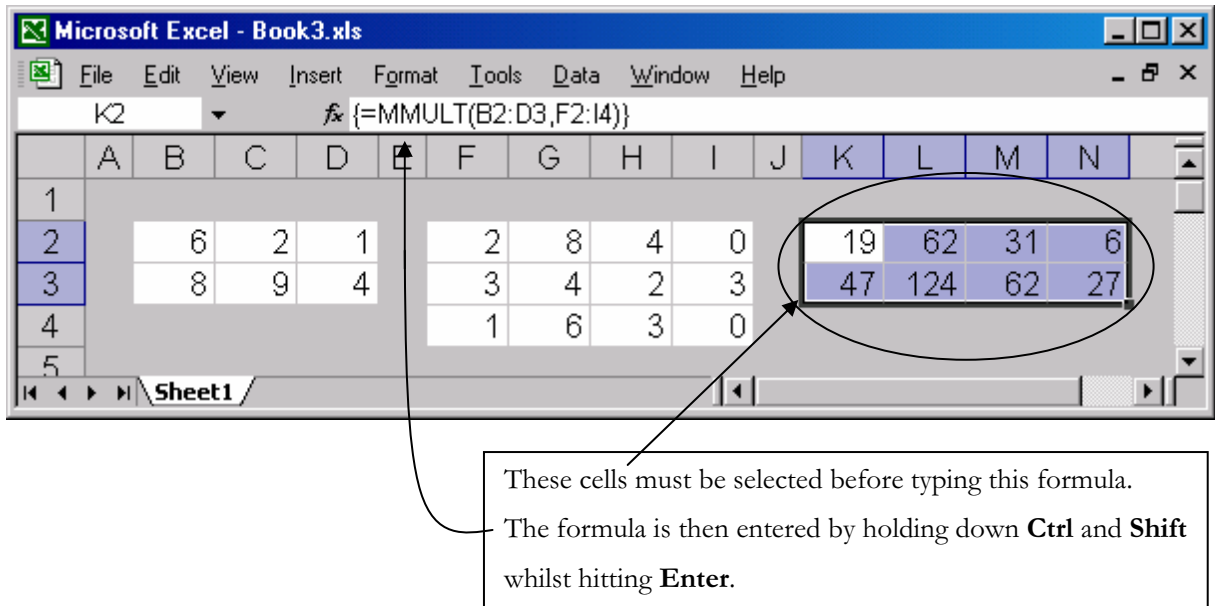
However, *matrix multiplication is not commutative*, that is, in general,

$$\mathbf{AB} \neq \mathbf{BA}.$$

Matrix Multiplication Using Excel

Excel has a function called MMULT for multiplying matrices (see Figure II.D.3). As with the addition of matrices, it is essential that the data are entered into the formula as an array. This means selecting the correct array of cells for the result, typing the formula, and then entering it by holding down **Ctrl** and **Shift** whilst hitting **Enter**.

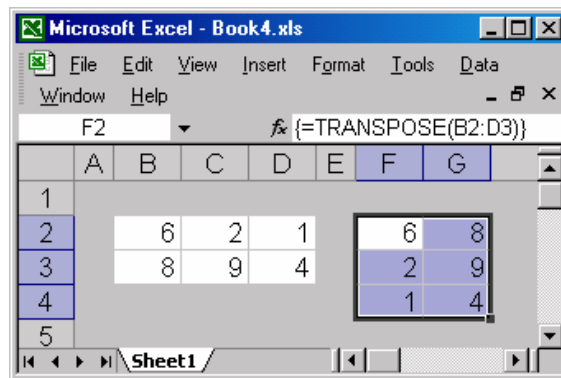
Figure II.D.3: Matrix multiplication in Excel



Transposing a Matrix in Excel

When transposing a matrix the top row of the original matrix becomes the first column of the transposed matrix, and so on (see Section II.D.1.2). Excel has a function called TRANSPOSE to carry out this process (see Figure II.D.4). Again, it requires the use of **Ctrl/Shift** and **Enter**.

Figure II.D.4: Matrix transposition in Excel



II.D.1.5 Inverting a Matrix

Matrix algebra can be used in the solving of simultaneous equations. This involves finding the *inverse* of a *square matrix*. To invert a matrix by hand is a tedious task, but fortunately packages such as Excel have a function to do it for us. The notation for the inverse of matrix **A** is \mathbf{A}^{-1} .

Only square matrices, and then only *some* square matrices, have inverses. The inverse of a matrix is a matrix such that the product of the matrix and its inverse is an identity matrix. An identity matrix is one in which every entry is zero except for the main diagonal which contains ones. Algebraically,

$$\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{ which is the } 2 \times 2 \text{ identity matrix, denoted } \mathbf{I}.$$

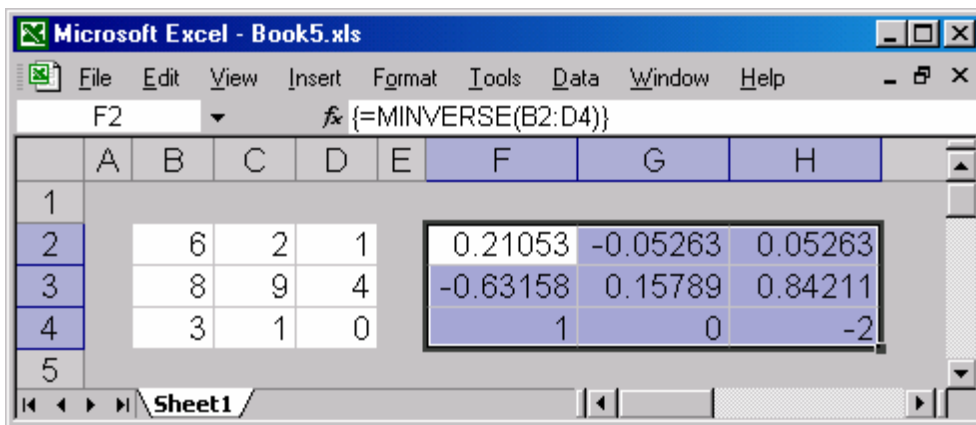
The identity matrix has the property that when used in matrix multiplication it leaves the multiplied matrix unchanged.

Matrices cannot be divided. However, if a square matrix has an inverse then we can instead multiply by that inverse to achieve the desired results (see below for an application). To derive the inverse of a matrix by hand, a partitioned matrix may be formed. This is achieved by positioning an identity matrix next to the matrix to be inverted, for example,

$$\left[\begin{array}{ccc|ccc} 6 & 2 & 1 & 1 & 0 & 0 \\ 8 & 9 & 4 & 0 & 1 & 0 \\ 3 & 1 & 0 & 0 & 0 & 1 \end{array} \right].$$

The objective is then to convert the original matrix into an identity matrix by adding or subtracting, multiples of rows together. When the matrix on the left is an identity matrix, the resultant matrix on the right will be the inverse of the original matrix. We will not illustrate this process here, instead we illustrate the use of the Excel function MINVERSE in Figure II.D.5. However, interested readers can see the process illustrated in Watsham and Parramore (1997).

Figure II.D.5: [Matrix inversion in Excel](#)



Matrix inversion underpins many statistical analyses as well as mathematical applications. For instance, statistical packages implement regression analysis (see Chapter II.F) by inverting a matrix.

II.D.2 Application of Matrix Algebra to Portfolio Construction

Matrix algebra can be applied in portfolio construction. We will illustrate this in two ways. The first is to calculate the risk for a portfolio where the asset weights are already known. The second is to solve a series of simultaneous equations to determine the asset weights that comprise a risk-minimising portfolio. Note that in this section we are drawing upon Chapter II.B and looking at historical returns.³ We are considering what would have happened to portfolios constructed using different weights. In the Chapter on probability we will be using the same formulae in the context of the expected value and variance of a random variable – that is, what might happen to differently weighted portfolios in the future. We will follow the convention of using Latin letters for statistics (e.g. s and r). Greek letters (σ and ρ) are used for the corresponding parameters (see Chapter II.E).

II.D.2.1 Calculating the Risk of an Existing Portfolio

In the case of a two-asset portfolio, if R_1 and R_2 are the asset returns to the assets, w is the portfolio weight on asset 1, that is, the proportion of the total value of the portfolio invested in asset 1, and R_p is the return to the portfolio, then

$$R_p = wR_1 + (1 - w)R_2. \quad (\text{see Section II.B.6.6})$$

(Note that w and $(1 - w)$ are known as the weights of the assets in the portfolio.)

In more complex cases we will be using vector notation for the return: $\mathbf{R} = \begin{bmatrix} R_1 \\ R_2 \end{bmatrix}$.

We saw in Chapter II.B that the risk in a portfolio is measured by the standard deviation of the returns to the portfolio:

$$\text{StDev}(R_p) = \sqrt{\text{Var}(R_p)},$$

and in the case of a two-asset portfolio

$$\text{Var}(R_p) = w^2 \text{Var}(R_1) + 2w(1 - w) \text{Covar}(R_1, R_2) + (1 - w)^2 \text{Var}(R_2). \quad (\text{II.D.2})$$

To facilitate generalisation later we recast this as:

$$\text{Var}(R_p) = w_1^2 \text{Var}(R_1) + 2w_1 w_2 \text{Covar}(R_1, R_2) + w_2^2 \text{Var}(R_2), \text{ where } w_1 + w_2 = 1. \quad (\text{II.D.3})$$

³ The second application also draws upon our knowledge of differential calculus.

We can use matrix algebra to write such expressions more easily, and also to evaluate them in Excel easily. We first need to introduce the concept of a variance–covariance matrix. For two assets this is:

$$\begin{bmatrix} \text{Var}(R_1) & \text{Covar}(R_1, R_2) \\ \text{Covar}(R_2, R_1) & \text{Var}(R_2) \end{bmatrix}$$

Following usual notation we shorten this to $\begin{bmatrix} s_1^2 & s_{12} \\ s_{21} & s_2^2 \end{bmatrix}$, remembering that $s_{21} = s_{12}$. These are (known) sample variances and covariances. Multiplying out the left-hand side below, we can show that⁴

$$\begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} s_1^2 & s_{12} \\ s_{21} & s_2^2 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \text{Var}(R_P) = w_1^2 s_1^2 + 2w_1 w_2 s_{12} + w_2^2 s_2^2$$

With three assets⁵

$$\text{Var}(R_P) = \underbrace{\begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix}}_{\mathbf{w}^T} \underbrace{\begin{bmatrix} s_1^2 & s_{12} & s_{13} \\ s_{21} & s_2^2 & s_{13} \\ s_{31} & s_{32} & s_3^2 \end{bmatrix}}_{\mathbf{V}} \underbrace{\begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}}_{\mathbf{w}}$$

and in general, with n assets in the portfolio:

$$\text{Var}(R_P) = \mathbf{w}^T \mathbf{V} \mathbf{w}, \tag{II.D.4}$$

where \mathbf{w} is the $n \times 1$ vector of portfolio weights⁶ and \mathbf{V} is the *covariance matrix* (or variance–covariance matrix) of the asset returns. The entry in the i th row and i th column of a covariance

⁴ To show this note that

$$\begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} s_1^2 & s_{12} \\ s_{21} & s_2^2 \end{bmatrix} = \begin{bmatrix} a & b \end{bmatrix}$$

where $a = w_1 s_1^2 + w_2 s_{21}$ and $b = w_1 s_{12} + w_2 s_2^2$, and

$$\begin{aligned} \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} &= a w_1 + b w_2 = (w_1 s_1^2 + w_2 s_{21}) w_1 + (w_1 s_{12} + w_2 s_2^2) w_2 = w_1^2 s_1^2 + w_1 w_2 s_{21} + w_1 w_2 s_{12} + w_2^2 s_2^2 \\ &= w_1^2 s_1^2 + 2w_1 w_2 s_{12} + w_2^2 s_2^2. \end{aligned}$$

⁵ Note that $\mathbf{w}^T \mathbf{V} \mathbf{w}$ has order given by $(1 \times 3) \times (3 \times 3) \times (3 \times 1)$. If we multiply this out, we obtain an order 1×1 matrix, that is, a *number*. (But remember that Excel distinguishes between a number and a 1×1 matrix.)

⁶ Recall that the transpose \mathbf{w}^T is the row vector obtained from \mathbf{w} by changing the column to a row.

matrix is the variance of the returns to the i th asset. The entry in the i th row and the j th column ($i \neq j$) is the covariance between the returns to the i th asset and the returns to the j th asset.

Example II.D.3:

Consider a three asset portfolio with the following details: $s_1^2 = 0.01887$, $s_2^2 = 0.02763$, $s_3^2 = 0.00966$, $s_{12} = 0.01722$, $s_{23} = 0.00604$, $s_{13} = 0.00721$, $w_1 = 0.04$, $w_2 = 0.03$, $w_3 = 0.03$. Calculate the volatility of the portfolio using matrix algebra.

The easiest way to calculate the variance of the portfolio is to first determine the weight matrix as well as the variance–covariance matrix, and then to use a program such as Excel to multiply the matrices. The task is to basically fill in the correct numbers in the right place in the individual matrices. Thus,

$$\text{Var}(R_p) = \begin{bmatrix} 0.4 & 0.3 & 0.3 \end{bmatrix} \begin{bmatrix} 0.01887 & 0.01722 & 0.00721 \\ 0.01722 & 0.02763 & 0.00604 \\ 0.00721 & 0.00604 & 0.00966 \end{bmatrix} \begin{bmatrix} 0.4 \\ 0.3 \\ 0.3 \end{bmatrix}.$$

Before showing how to calculate this in Excel, we wish to demonstrate how it can be done manually. Multiplying the rows of the weightings vector by the columns of the variance–covariance matrix gives one element of a new matrix. In this example it will be a row vector, $\mathbf{w}^T\mathbf{V}$. This is illustrated below.

(0.4×0.01887)	(0.4×0.01722)	(0.4×0.00721)
$+ (0.3 \times 0.01722)$	$+ (0.3 \times 0.02763)$	$+ (0.3 \times 0.00604)$
$+ (0.3 \times 0.00721)$	$+ (0.3 \times 0.00604)$	$+ (0.3 \times 0.00966)$
$= 0.007548 + 0.005166 + 0.002163$	$= 0.006888 + 0.008289 + 0.001812$	$= 0.002884 + 0.001812 + 0.002898$
$= 0.0014877$	$= 0.016989$	$= 0.007594$

Therefore, the $\mathbf{w}^T\mathbf{V}$ vector is equal to

$$\begin{bmatrix} 0.014877 & 0.016989 & 0.007594 \end{bmatrix}.$$

We now post-multiply the $\mathbf{w}^T\mathbf{V}$ vector by the column vector of weightings giving

$$\begin{bmatrix} 0.014877 & 0.016989 & 0.007594 \end{bmatrix} \begin{bmatrix} 0.4 \\ 0.3 \\ 0.3 \end{bmatrix},$$

which is

$$\begin{aligned} (0.014877 \times 0.4) + (0.016989 \times 0.3) + (0.007594 \times 0.3) &= 0.005951 + 0.005097 + 0.002279 \\ &= 0.01333. \end{aligned}$$

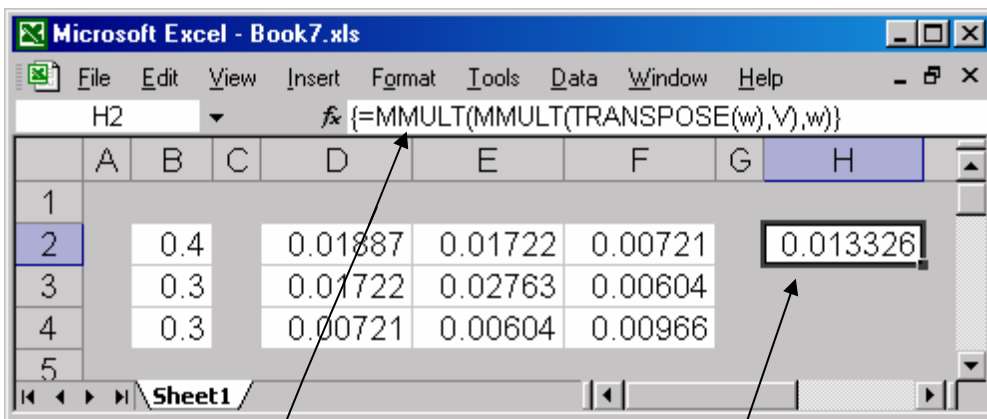
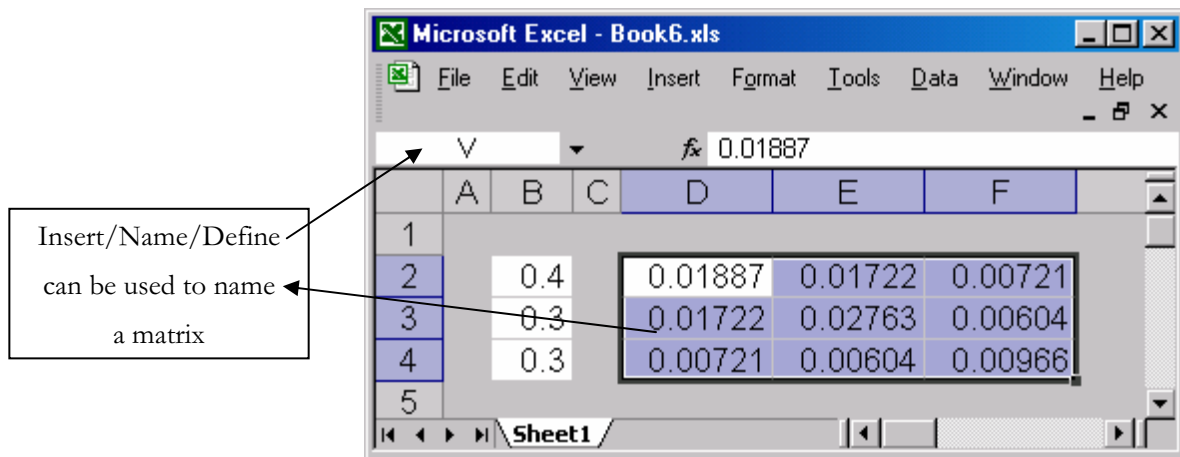
This is the variance. To find the standard deviation we take the square root of this:

$$\sqrt{0.0133} = 0.1155 = 11.55\%$$

Therefore, the standard deviation of the returns to the portfolio is 11.55%. If the asset returns are annual returns, the standard deviation of portfolio is already in annual terms. This is its *volatility*.

As previously stated, the variance of a portfolio can be calculated using just one formula in Excel. Figure II.D.6 shows you how to do this, using the previous example to illustrate.

Figure II.D.6: **Calculating the variance of a portfolio in Excel**



But note:
Excel regards the result as a 1×1 matrix, so the formula must be entered using Ctrl/Shift and Enter

II.D.2.2 Deriving Asset Weights for the Minimum Risk Portfolio

Suppose now that we have three assets with return variances and covariances as in Example II.D.3 above, and suppose in addition that their historical returns over the period of interest were 8%, 10% and 7% respectively. Over the period in question what portfolio would have delivered a return of, say, 9% with minimum risk? This is the *Markowitz* problem (see Markowitz, 1952). In this case we want to find the portfolio weights w_1 , w_2 and w_3 to minimize the variance subject to $\text{sumproduct}(\mathbf{w}, [0.08 \ 0.10 \ 0.07]) = 0.09$ (and, of course, that the weights add up to one).

The general Markowitz problem is to find weights \mathbf{w} such that:

$$\mathbf{w}^T \mathbf{V} \mathbf{w} \text{ is minimized subject to } w_1 + \dots + w_n = 1 \text{ and } \mathbf{w}^T \mathbf{R} = r. \quad (\text{II.D.5})$$

This is a constrained optimisation problem for which we need the techniques developed in Chapter II.C. Specifically, we need the *Lagrangian* (Section II.C.7.4).

Example II.D.4:

Writing out the problem in full, with the constraints in "=0" form, we get:

$$\begin{aligned} \text{Minimise} \quad & w_1^2 s_1^2 + w_2^2 s_2^2 + w_3^2 s_3^2 + 2w_1 w_2 s_{12} + 2w_1 w_3 s_{13} + 2w_2 w_3 s_{23} \\ \text{subject to} \quad & w_1 + w_2 + w_3 - 1 = 0 \\ \text{and} \quad & 0.08w_1 + 0.10w_2 + 0.07w_3 - 0.09 = 0. \end{aligned}$$

(We avoid substituting in the values for the standard deviations for the moment – in the interests of clarity!) The Lagrangian is

$$\begin{aligned} L = & w_1^2 s_1^2 + w_2^2 s_2^2 + w_3^2 s_3^2 + 2w_1 w_2 s_{12} + 2w_1 w_3 s_{13} + 2w_2 w_3 s_{23} \\ & + \lambda_1(w_1 + w_2 + w_3 - 1) + \lambda_2(0.08w_1 + 0.10w_2 + 0.07w_3 - 0.09). \end{aligned}$$

Recall that the solution to the constrained optimization problem is given by finding the minimum of the Lagrangian, and that, for this, we require that all five partial derivatives be zero. This gives the following five equations:

$$\left(\frac{\partial L}{\partial w_1} = \right) 2w_1 s_1^2 + 2w_2 s_{12} + 2w_3 s_{13} + \lambda_1 + 0.08\lambda_2 = 0$$

$$\left(\frac{\partial L}{\partial w_2} = \right) 2w_2 s_2^2 + 2w_1 s_{12} + 2w_3 s_{23} + \lambda_1 + 0.10\lambda_2 = 0$$

$$\left(\frac{\partial L}{\partial w_3} = \right) 2w_3 s_3^2 + 2w_1 s_{13} + 2w_2 s_{23} + \lambda_1 + 0.07\lambda_2 = 0$$

$$\left(\frac{\partial L}{\partial \lambda_1} = \right) w_1 + w_2 + w_3 - 1 = 0$$

$$\left(\frac{\partial L}{\partial \lambda_2} = \right) 0.08w_1 + 0.10w_2 + 0.07w_3 - 0.09 = 0.$$

Rewriting in order $w_1, w_2, w_3, \lambda_1, \lambda_2$, etc.:

$$2s_1^2 w_1 + 2s_{12} w_2 + 2s_{13} w_3 + \lambda_1 + 0.08 \lambda_2 = 0$$

$$2s_{12} w_1 + 2s_2^2 w_2 + 2s_{23} w_3 + \lambda_1 + 0.10 \lambda_2 = 0$$

$$2s_{13} w_1 + 2s_{23} w_2 + 2s_3^2 w_3 + \lambda_1 + 0.07 \lambda_2 = 0$$

$$w_1 + w_2 + w_3 + 0\lambda_1 + 0\lambda_2 = 1$$

$$0.08 w_1 + 0.10 w_2 + 0.07 w_3 + 0\lambda_1 + 0\lambda_2 = 0.09 .$$

Such a system of simultaneous linear equations can be written out in matrix form by extracting the

matrix of coefficients, $\begin{bmatrix} 2s_1^2 & 2s_{12} & 2s_{13} & 1 & 0.08 \\ 2s_{12} & 2s_2^2 & 2s_{23} & 1 & 0.10 \\ 2s_{13} & 2s_{23} & 2s_3^2 & 1 & 0.07 \\ 1 & 1 & 1 & 0 & 0 \\ 0.08 & 0.10 & 0.07 & 0 & 0 \end{bmatrix}$, the vector of variables, $\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \lambda_1 \\ \lambda_2 \end{bmatrix}$, and the vector

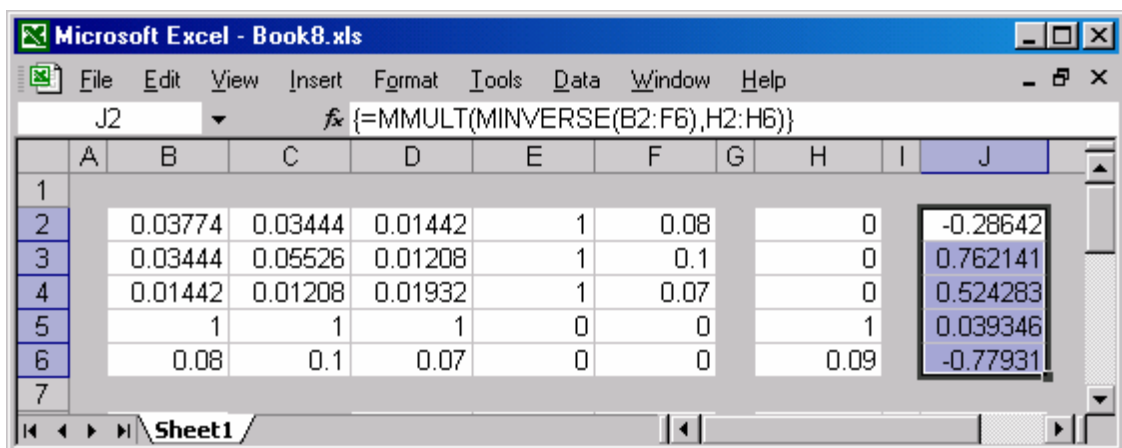
of right-hand sides, $\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0.09 \end{bmatrix}$. The system can then be written

$$\begin{bmatrix} 2s_1^2 & 2s_{12} & 2s_{13} & 1 & 0.08 \\ 2s_{12} & 2s_2^2 & 2s_{23} & 1 & 0.10 \\ 2s_{13} & 2s_{23} & 2s_3^2 & 1 & 0.07 \\ 1 & 1 & 1 & 0 & 0 \\ 0.08 & 0.10 & 0.07 & 0 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0.09 \end{bmatrix} .$$

Solution

Figure II.D.7 shows the result of substituting the known numerical values for the s 's in Excel and solving, using the matrix inverse.

Figure II.D.7: Solution to Example II.D.4

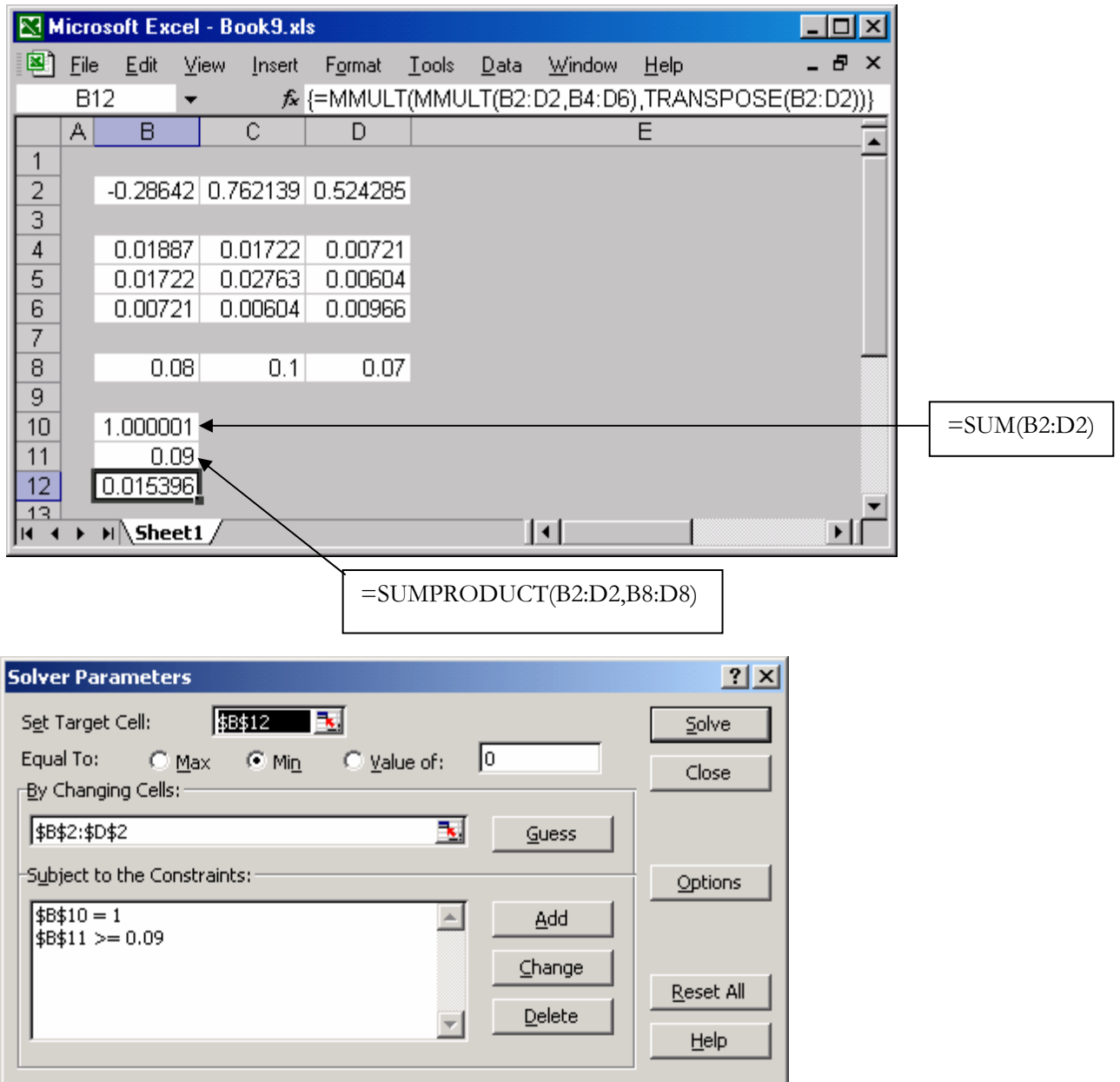


Thus the optimal portfolio is given by $w_1 = -0.286$, $w_2 = 0.762$ and $w_3 = 0.524$. These *do* sum to 1, the negative value indicating that asset 1 has been shorted. If this is not acceptable then the problem must be reformulated with an extra constraint, $w_1 \geq 0$, or if none of the assets can be shorted, $w_1 \geq 0$ and $w_2 \geq 0$ and $w_3 \geq 0$. The values of the Lagrange multipliers have interpretations, but we shall not pursue that here.

Using Solver

Excel has a built-in constrained optimizer called Solver (see Figure II.D.8). This will do most of this work automatically. It is an add-in under the Tools menu. (Your MS Office disk may be needed to add it in, depending on how Office was installed on your machine.)

Figure II.D.8: [Using Solver in Example II.D.4](#)



II.D.2.3 Hedging a Vanilla Option Position

In Chapter I.A.8 it is shown that vanilla equity option prices are a function of variables such as the asset price, the exercise (or strike) price, the life of the option, the risk-free rate of interest, the volatility of the underlying asset and the dividend rate (or equivalent cash distribution). In Sections II.C.4.3 and II.C.5.1 we saw that:

- the first derivative of the option price with respect to the price of the underlying is known as *delta*;
- the second derivative of the option price with respect to the price of the underlying is known as *gamma*;
- the first derivative of the option price with respect to volatility is known as *vega*.

Also:

- the first derivative of the option price with respect to the eroding life of the option is known as *theta*;
- the first derivative of the option price with respect to the risk-free rate of interest is known as *rho*.

Habitually, option market makers will hedge their options book against the effects of changes in the price of the underlying (delta and gamma) and changes in volatility (vega) by ensuring that their options book is delta-, gamma- and vega-neutral. The first stage of any strategy to hedge an option position is to calculate the position delta and then create a hedge so that the position is *delta-neutral*.

This requires two processes:

- 1) Calculation of the *position delta*.
- 2) Establishing the delta-neutral hedge.

II.D.2.3.1 Calculating the position delta

Because differentiation is a linear process, deltas (and the other ‘Greeks’) combine linearly. Consequently, the delta of a portfolio of options, Δ_{Π} , is the value-weighted average of the deltas of each of the options:

$$\Delta_{\Pi} = \sum_{i=1}^n w_i \Delta_i, \quad (\text{II.D.6})$$

where w_i is the proportion of the value of the portfolio accounted for by the value of the position in option i , and Δ_i is the delta of option i .

Each position is in options on the same underlying asset. However, the position can include puts and calls, long and short positions, and options with different exercise prices and expiry dates. Although delta is usually considered as a concept associated with options, it is, as we noted above,

no more than the first derivative of the option price with respect to the asset price. Similarly, the delta of the underlying, i.e. the first derivative of the asset price with respect to itself, is 1.

II.D.2.3.2 Establishing the delta-neutral hedge

If the option position had a net delta of +0.6, that option position could be made delta-neutral by going short the underlying asset to a value equal to 60% of the value of the notional assets underlying the options position. For example, if the option position consisted of calls on 100,000 shares, going short 60,000 of those shares will make the option position instantaneously delta-neutral.

To illustrate how this works, assume that the underlying asset falls by one unit (and delta remains constant!). Then each option will fall in price by 0.6 and the long position in the option will lose 60,000. A short position in 60,000 units of the underlying asset will show a profit of 60,000 if the underlying asset falls in price by 1, thereby offsetting the loss on the options.

However, the word ‘instantaneously’ appears because delta changes through time and as the asset price changes. In the Black–Scholes world of continuous diffusion, known and constant volatility and no transactions costs, it would be possible to continuously rebalance the portfolio by changing the short position in the underlying in order to ensure that the portfolio remains delta-neutral. However, in reality there is no continuous market, there are transactions costs and we do not know the volatility, but we do know that it is not constant.

If the portfolio cannot be rebalanced continuously so that it is continuously delta-neutral, the gamma of the option becomes relevant. (Arguably the theta will also be relevant, but empirically this is less significant.) If the implied volatility in the option is not constant then vega is also important.

The relative influences of delta, gamma and vega can be seen in the following equation (cf. equation (II.C.14)):

$$\partial W \approx \Delta \partial S + V \partial \sigma + \frac{1}{2} \Gamma \partial S^2 \quad (\text{II.D.7})$$

where ∂W is the change in value of the option, S is the value of the underlying asset, ∂S is the change in value of the underlying, and $\partial \sigma$ the change in its volatility. Δ is the delta of the option portfolio, V its vega and Γ its gamma.

II.D.2.3.3 Gamma neutrality

Although an option position can be made delta-neutral by trading in the underlying asset, such a trade will not manage the gamma risk. This is because only options exhibit gamma risk. Long option positions will have positive gamma.

If we let the gamma of the delta-neutral position be Γ_χ and the gamma of any additional options be Γ_h , then the number of options to sell in order to make the position gamma-neutral is given by the ratio $-\Gamma_\chi/\Gamma_h$ multiplied by the number of options in the long position. However, the options sold will themselves have a delta and therefore the position will no longer be delta-neutral. So a trade in the underlying asset must be executed to re-establish delta neutrality.

II.D.2.3.4 Vega neutrality

As only options exhibit vega risk, again it is necessary to use options in order to hedge the vega risk in the option position. The number of a given option to trade in order to establish vega neutrality is given by the original position multiplied by the ratio $-V_\chi/V_h$.

II.D.2.3.5 Hedging a short option position

We will illustrate the hedging of a short position of 100,000 options with an exercise price of 150, a life of 3 months, an underlying asset price of 150 and a volatility of 25%. There are no dividends. The risk-free rate of interest is 6%. The option premium is 8.5895. The delta of the option is 0.5724, the gamma is 0.0209 and the vega is 29.4265.

In order to be delta-, gamma- and vega-neutral it is necessary to have at least two other different types of options in the hedging strategy. We will assume that the first option to be used for hedging is a 160 call. For the second we shall use the 170 call.

Calls	Price	Delta	Gamma	Vega
150(ATM)	08.58953	0.57240	0.02093	29.42652
160	04.44943	0.36926	0.02012	28.29926
170	02.05636	0.20645	0.01522	21.39867

In order to achieve delta, gamma and vega neutrality simultaneously, we need to solve the following simultaneous equations for x_1 and x_2 which represent the numbers of 160 and 170 calls respectively:

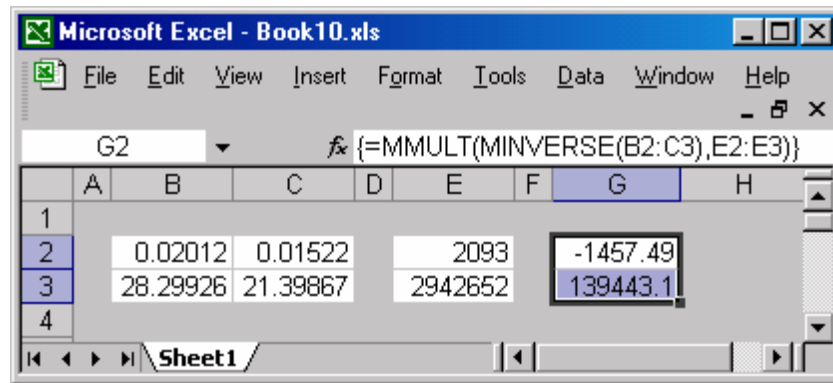
$$\begin{aligned} \Gamma_{\Pi} + \Gamma_1 x_1 + \Gamma_2 x_2 &= 0, \\ V_{\Pi} + V_1 x_1 + V_2 x_2 &= 0. \end{aligned} \tag{II.D.8}$$

The short position in the 150 calls had a position gamma of -2093 and a position vega of $-2,942,652$. So to achieve gamma and vega neutrality we must solve the following equations:

$$\begin{aligned} -2093 + 0.02012x_1 + 0.01522x_2 &= 0 \\ -2942652 + 28.29926x_1 + 21.39867x_2 &= 0 \end{aligned} \tag{II.D.9}$$

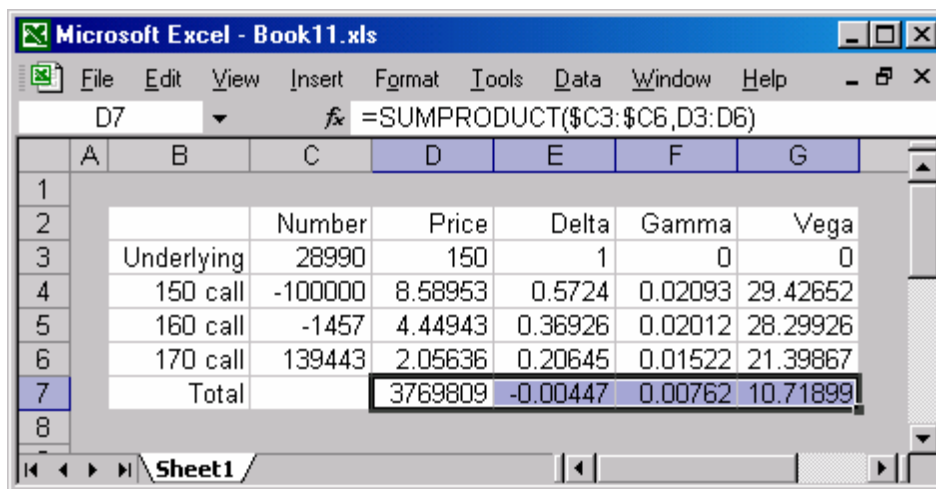
We can use matrix algebra, implemented in Excel, to solve these equations (Figure II.D.9; see Section II.D.1.5 above).

Figure II.D.9: Solving equations (II.D.9) in Excel



The solutions are to go short 1457 of the 160 calls and long 139,443 of the 170 calls. The deltas of these option positions are -538 and $28,788$, respectively. Thus combining with the delta of the short position in the 150 calls of $-57,240$, the net delta of the three option positions is $-28,990$ (Figure II.D.10). Thus 28,990 units of the underlying asset will have to be purchased to make the position delta-neutral.

Figure II.D.10: Calculating hedge ratios for a portfolio in Excel



II.D.3 Quadratic Forms

A quadratic expression in two variables, x and y , is an algebraic expression involving x^2 , y^2 and xy terms. Similarly, a quadratic expression in three variables, x , y and z , involves x^2 , y^2 , z^2 , xy , xz and yz terms. For example, the variance of a portfolio of three assets A, B and C involves terms in w_1^2 , w_2^2 , w_3^2 , w_1w_2 , w_1w_3 and w_2w_3 , namely,

$$s_1^2w_1^2 + s_2^2w_2^2 + s_3^2w_3^2 + 2s_{12}w_1w_2 + 2s_{13}w_1w_3 + 2s_{23}w_2w_3.$$

(Note that we are regarding the standard deviations as known numbers.)

We have seen that this can be expressed in matrix form as (II.D.4):

$$\begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix} \begin{bmatrix} s_1^2 & s_{12} & s_{13} \\ s_{21} & s_2^2 & s_{13} \\ s_{31} & s_{32} & s_3^2 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}.$$

All quadratic forms can be expressed similarly. The matrix is 0.5 times the matrix of second partial derivatives. The matrix of second partial derivatives is called the *Hessian* matrix.

Thus if $f(x,y) = 4x^2 + 6xy - 3y^2$, then the Hessian is

$$\begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} = \begin{bmatrix} 8 & 6 \\ 6 & -6 \end{bmatrix},$$

and

$$4x^2 + 6xy - 3y^2 = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 4 & 3 \\ 3 & -3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

This generalises easily to more than two variables. For example, if

$$g(x,y,z) = 4x^2 - 3y^2 + 5z^2 + 6xy - 2xz + 12yz$$

then the Hessian is

$$\begin{bmatrix} 8 & 6 & -2 \\ 6 & -6 & 12 \\ -2 & 12 & 10 \end{bmatrix}$$

and

$$g(x,y,z) = \begin{bmatrix} x & y & z \end{bmatrix} \begin{bmatrix} 4 & 3 & -1 \\ 3 & -3 & 6 \\ -1 & 6 & 5 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}.$$

II.D.3.1 The Variance of Portfolio Returns as a Quadratic Form

We have already seen the importance of one particular quadratic form – that which gives the variance of the return to a portfolio. This is sometimes rewritten using the relationship between standard deviations (e.g., s_1 and s_2), the covariance (s_{12}), and the corresponding correlation coefficient (r_{12}). [Reminder: We are following the convention of using Latin letters for statistics (e.g., s and r). Greek letters (σ and ρ) are used for the corresponding parameters.] That relationship may be found in Chapters II.B and II.E and is:

$$s_{12} = s_1 s_2 r_{12}. \quad (\text{II.D.10})$$

We can generalise (II.D.10) to the matrix form:

$$\mathbf{V} = \mathbf{DCD} \quad (\text{II.D.11})$$

which, in the case of three securities, has the following form:

$$\begin{bmatrix} s_1^2 & s_{12} & s_{13} \\ s_{21} & s_2^2 & s_{23} \\ s_{31} & s_{32} & s_3^2 \end{bmatrix} = \begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & s_3 \end{bmatrix} \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix} \begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & s_3 \end{bmatrix} \quad (\text{II.D.12})$$

\mathbf{C} is known as the *correlation matrix*, and \mathbf{D} is so labelled since it is an example of a diagonal matrix.

The variance of the returns to a three-asset portfolio,

$$\begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix} \begin{bmatrix} s_1^2 & s_{12} & s_{13} \\ s_{21} & s_2^2 & s_{23} \\ s_{31} & s_{32} & s_3^2 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix},$$

can thus be written:

$$\begin{aligned} & \begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix} \begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & s_3 \end{bmatrix} \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix} \begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & s_3 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \\ & = \begin{bmatrix} w_1 s_1 & w_2 s_2 & w_3 s_3 \end{bmatrix} \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix} \begin{bmatrix} w_1 s_1 \\ w_2 s_2 \\ w_3 s_3 \end{bmatrix}. \end{aligned} \quad (\text{II.D.13})$$

II.D.3.2 Definition of Positive Definiteness

If a matrix \mathbf{M} has the property that for all non-zero vectors \mathbf{w} , $\mathbf{w}^T\mathbf{M}\mathbf{w} > 0$, then \mathbf{M} is said to be *positive definite*.

If a matrix \mathbf{M} has the property that for all non-zero vectors \mathbf{w} , $\mathbf{w}^T\mathbf{M}\mathbf{w} < 0$, then \mathbf{M} is said to be *negative definite*.

If it has the property that for all non-zero vectors \mathbf{w} , $\mathbf{w}^T\mathbf{M}\mathbf{w} \geq 0$, then \mathbf{M} is said to be *positive semi-definite*.

If it has the property that for all non-zero vectors \mathbf{w} , $\mathbf{w}^T\mathbf{M}\mathbf{w} \leq 0$, then \mathbf{M} is said to be *negative semi-definite*.

These considerations are important in *optimisation* (see Section II.C.7). It is also important to note that a covariance matrix should be positive definite (see also Section II.D.5.4.1).

II.D.4 Cholesky Decomposition

There are many situations in which we may need to simulate the way in which a portfolio of assets behaves, notably in working on value-at-risk (VaR), but also when valuing some exotic options. Essentially we need to simulate a vector of returns. To do this we need to capture the structure of the covariance matrix. This can be achieved by splitting the matrix \mathbf{V} into the product of a lower triangular matrix \mathbf{L} and an upper triangular matrix \mathbf{U} , so that $\mathbf{V} = \mathbf{L}\mathbf{U}$. (It turns out that it is the positive definiteness of \mathbf{V} which ensures that this can be achieved, but we shall pursue that no further.)

In the three-asset case:

$$\mathbf{V} = \begin{bmatrix} \text{Var}(R_1) & \text{Covar}(R_1, R_2) & \text{Covar}(R_1, R_3) \\ \text{Covar}(R_2, R_1) & \text{Var}(R_2) & \text{Covar}(R_2, R_3) \\ \text{Covar}(R_3, R_1) & \text{Covar}(R_3, R_2) & \text{Var}(R_3) \end{bmatrix} = \begin{bmatrix} a & 0 & 0 \\ x & b & 0 \\ y & z & c \end{bmatrix} \begin{bmatrix} a & x & y \\ 0 & b & z \\ 0 & 0 & c \end{bmatrix} = \mathbf{L}\mathbf{U}.$$

The quantities a, b, c, x, y and z have to be determined by some simple arithmetic (see below). We can show that if

$$\mathbf{n} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

is a vector of independent standard normal random variables then \mathbf{Lz} has covariance matrix \mathbf{V} . This is because $\mathbf{V}(\mathbf{Lz}) = \mathbf{L}\mathbf{V}(\mathbf{z})\mathbf{L}^T$, where $\mathbf{V}(\mathbf{x})$ stands for ‘the covariance matrix of the elements of \mathbf{x} ’. But

$$\mathbf{V}(\mathbf{z}) = \mathbf{V} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

since the z s are independent and have variances of 1. So:

$$\mathbf{V}(\mathbf{Lz}) = \mathbf{L} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{L}^T = \mathbf{L}\mathbf{L}^T = \mathbf{L}\mathbf{U} = \mathbf{V}.$$

Furthermore, the components of \mathbf{Lz} are just linear combinations of z_1 , z_2 and z_3 , and linear combinations of normally distributed variables are normally distributed. So \mathbf{Lz} is a vector of normally distributed components with covariance matrix \mathbf{V} , which is exactly what we need.

II.D.4.1 The Cholesky Arithmetic

We demonstrate how to do the \mathbf{LU} decomposition in a three-asset example. Let

$$\mathbf{V} = \begin{bmatrix} 0.0144 & 0.0199 & 0.0148 \\ 0.0199 & 0.0529 & 0.0231 \\ 0.0148 & 0.0231 & 0.0225 \end{bmatrix}.$$

Now

$$\begin{bmatrix} a & 0 & 0 \\ x & b & 0 \\ y & z & c \end{bmatrix} \begin{bmatrix} a & x & y \\ 0 & b & z \\ 0 & 0 & c \end{bmatrix} = \begin{bmatrix} a^2 & ax & ay \\ ax & x^2 + b^2 & xy + bz \\ ay & xy + bz & y^2 + z^2 + c^2 \end{bmatrix},$$

so $a = \sqrt{0.0144} = 0.12$

$$x = \frac{0.0199}{a} = 0.1658333$$

$$y = \frac{0.0148}{a} = 0.1233333$$

$$b = \sqrt{0.0529 - x^2} = 0.1593716$$

$$z = \frac{0.0231 - xy}{b} = 0.0166104$$

$$c = \sqrt{0.0225 - y^2 - z^2} = 0.0837436$$

Hence

$$\mathbf{L} \approx \begin{bmatrix} 0.12 & 0 & 0 \\ 0.1658 & 0.1594 & 0 \\ 0.1233 & 0.0166 & 0.0837 \end{bmatrix}.$$

II.D.4.2 Simulation in Excel

To simulate the returns to (say) three assets:

1. estimate the annual returns and the returns covariance matrix;
2. decompose the covariance matrix into **L** and **U** (Cholesky);
3. use the **L** matrix to ‘convert’ a unit normal vector to give a simulated deviation from the expected return.

Example II.D.5:

Suppose

$$\begin{bmatrix} R_1 \\ R_2 \\ R_3 \end{bmatrix} = \begin{bmatrix} 0.07 \\ 0.12 \\ 0.10 \end{bmatrix} \text{ and } \mathbf{V} = \begin{bmatrix} 0.0144 & 0.0199 & 0.0148 \\ 0.0199 & 0.0529 & 0.0231 \\ 0.0148 & 0.0231 & 0.0225 \end{bmatrix}$$

(as above in Section II.D.4.1). Then

$$\mathbf{L} \approx \begin{bmatrix} 0.12 & 0 & 0 \\ 0.1658 & 0.1594 & 0 \\ 0.1233 & 0.0166 & 0.0837 \end{bmatrix}$$

(again as above). The final step is shown in Figure II.D.11.

Figure II.D.11: [Simulating correlated returns in Example II.D.5 in Excel](#)

Microsoft Excel - Book12.xls

File Edit View Insert Format Tools Data
Window Help

F4 fx =NORMSINV(RAND())

1					
2		0.07	0.12	0.1	
3					
4		0.1200	0.0000	0.0000	-0.09374
5		0.1658	0.1594	0.0000	2.009367
6		0.1233	0.0166	0.0837	0.533663
7					

Sheet1

This gives a realisation of a standard normal variable, that is, a normal variable with mean 0 and variance 1.

It is repeated in cells F5 and F6.

Microsoft Excel - Book13.xls

File Edit View Insert Format Tools Data Window Help

H4 fx =MMULT(B4:D6,F4:F6)

1					
2		0.07	0.12	0.1	
3					
4		0.1200	0.0000	0.0000	-1.04621
5		0.1658	0.1594	0.0000	0.266047
6		0.1233	0.0166	0.0837	1.409618
7					

Sheet1

This has now converted the independent normal variates into a random sample of deviations corresponding to our correlated and non-standard returns.

Remember – **Ctrl/Shift** and **Enter** are needed!

Microsoft Excel - Book14.xls

File Edit View Insert Format Tools Data Window Help

J4 fx =TRANSPOSE(B2:D2)+H4:H6

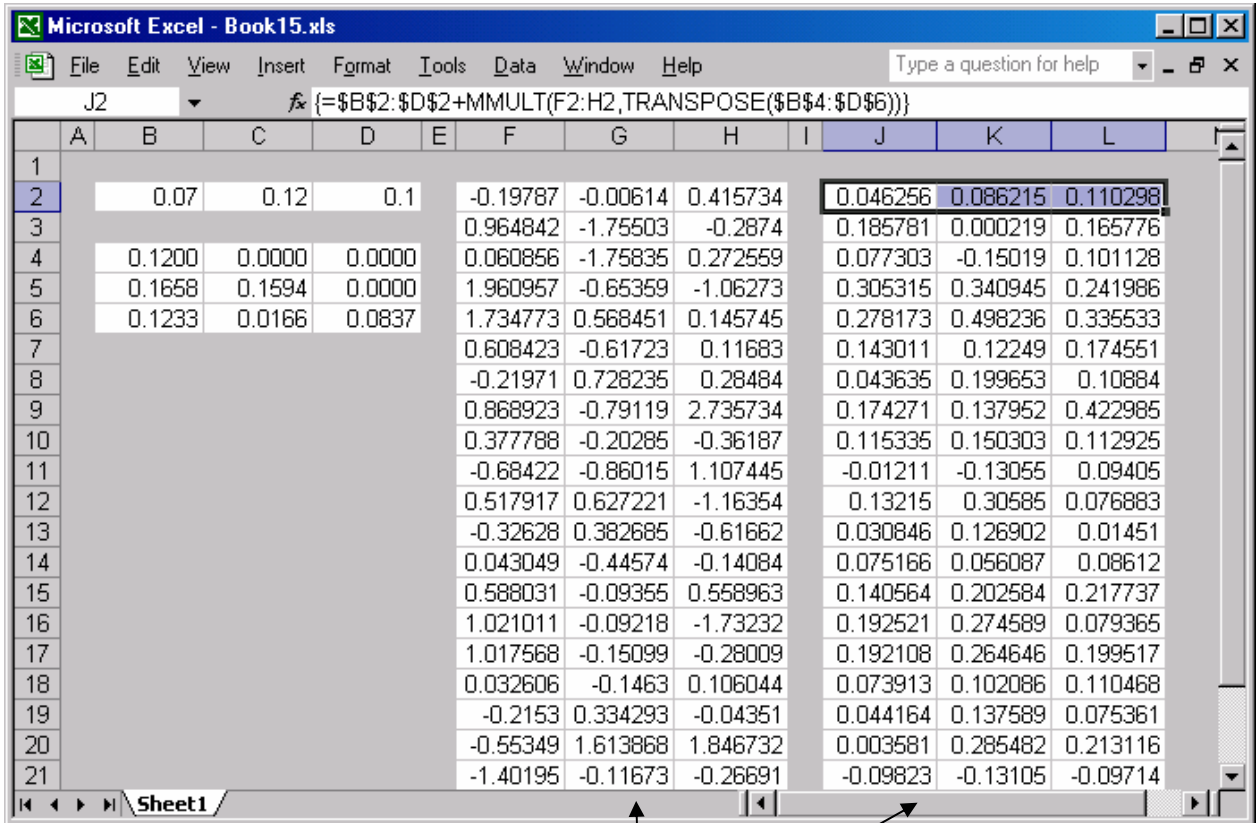
1						
2		0.07	0.12	0.1		
3						
4		0.1200	0.0000	0.0000	-1.79415	-0.2153
5		0.1658	0.1594	0.0000	-2.05009	-0.62425
6		0.1233	0.0166	0.0837	-0.31034	-0.28123
7						
8						

Sheet1

Finally we add on the vector of mean returns to give a simulated vector of returns. Note that this involves using the TRANSPOSE function.

Whilst Figure II.D.11 spreadsheets illustrate the approach, in practice the computations would be compressed to allow for many repetitions of the simulation. Those simulated returns can then be used in constructing VaR calculations, or in simulating the prices of underlying securities in option calculations. This is shown in Figure II.D.12

Figure II.D.12: [Simulating correlated returns in Excel](#)



We can copy these rows down to produce as many simulations as we require. The simulated returns in columns J, K and L will mimic the means and the variance–covariance construction of the observed returns.

The parameters can be adjusted to give simulated returns over different required time periods

If one is, say, trying to estimate the value of a portfolio by simulating the price changes of the underlying assets, then many repetitions will be needed. Just how many can be calculated, given the confidence required and the degree of accuracy required. It involves running a pilot simulation with, perhaps, a thousand repetitions, to estimate the standard deviation of the future worth of the option. The calculation involves the *standard error of the mean*; see Section II.G.3.2.

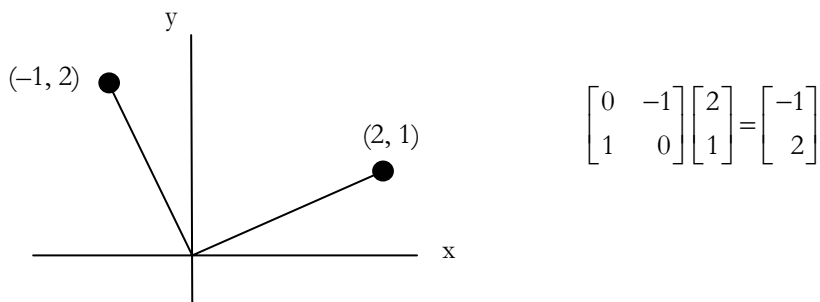
II.D.5 Eigenvalues and Eigenvectors

Finally, we examine an area of linear mathematics which has many applications, outlining where it is used in finance. It concerns square matrices. In general, a matrix may be regarded as a representing a transformation. A square matrix represents a transformation of space on to itself.

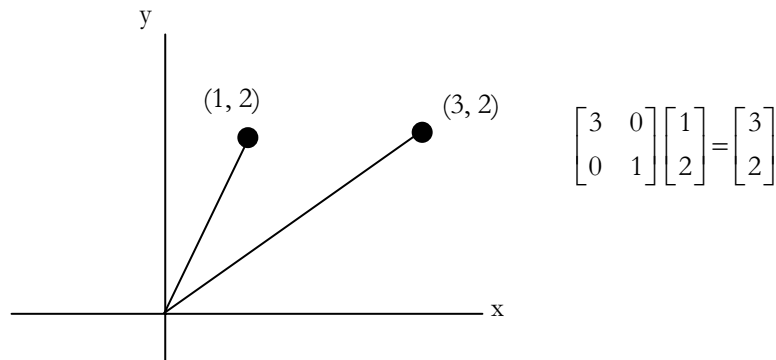
II.D.5.1 Matrices as Transformations

Square matrices can be viewed geometrically by considering their effect when pre-multiplying vectors. A 2×2 matrix can be considered as a transformation of the plane. Let us look at a couple of examples.

$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ represents an anticlockwise rotation of 90° about the origin:



$\begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$ represents a stretch by a factor of 3 in the x -direction:



(Note that $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ is always $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$.)

Example II.D.6:

How does the matrix $\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ translate the vectors $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$?

Doing the calculations, we have:

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix} = 3 \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} = -1 \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Thus the vectors $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ seem to be rather special vectors for this matrix. For almost any other choice of vector, the matrix would have translated them to another vector that was not just a multiple of that vector. For instance,

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ 3 \end{bmatrix} = \begin{bmatrix} 10 \\ 11 \end{bmatrix}$$

and $(10, 11)^T$ does not line in the same line as $(4, 3)^T$. However, $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ are special vectors

for $\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$, because $\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ translates them to other vectors which are on the same line.

In fact, $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ are the *eigenvectors* of $\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$.

What are their ‘eigenvalues?’ Guess....

$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ has eigenvalue 3 and $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ has eigenvalue -1 . We will now define these quantities formally.

II.D.5.2 Definition of Eigenvector and Eigenvalue

- (i) A square matrix \mathbf{A} has *eigenvector* \mathbf{x} if $\mathbf{Ax} = \lambda\mathbf{x}$.
- (ii) The constant λ is called the *eigenvalue* associated with \mathbf{x} .

Example II.D.7:

Consider the matrix $\mathbf{A} = \begin{bmatrix} 4.6 & -1.2 \\ -1.2 & 1.4 \end{bmatrix}$. Its eigenvalues can be found as follows:

$$\begin{bmatrix} 4.6 & -1.2 \\ -1.2 & 1.4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \end{bmatrix} \Rightarrow \begin{aligned} 4.6x - 1.2y &= \lambda x \\ -1.2x + 1.4y &= \lambda y \end{aligned}$$

so that

$$\begin{aligned}(4.6 - \lambda)x - 1.2y &= 0 \\ -1.2x + (1.4 - \lambda)y &= 0\end{aligned}$$

and thus

$$\frac{4.6 - \lambda}{1.2} = \frac{1.2}{1.4 - \lambda}.$$

If we cross-multiply and put all terms on the left-hand side

$$(4.6 - \lambda)(1.4 - \lambda) - 1.44 = 0, \text{ i.e. } \lambda^2 - 6\lambda + 5 = 0, \text{ i.e. } (\lambda - 5)(\lambda - 1) = 0.$$

So the eigenvalues are 5 and 1.

For the eigenvalue 5,

$$\begin{bmatrix} 4.6 & -1.2 \\ -1.2 & 1.4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 5 \begin{bmatrix} x \\ y \end{bmatrix} \text{ giving } \begin{aligned} 4.6x - 1.2y &= 5x, \\ -1.2x + 1.4y &= 5y. \end{aligned}$$

Both equations simplify to $y = -\frac{1}{3}x$. So *any* two-dimensional vector with

$$(y \text{ coordinate}) = -\frac{1}{3} \times (x \text{ coordinate})$$

is an eigenvector with eigenvalue 5. For example,

$$\begin{bmatrix} 4.6 & -1.2 \\ -1.2 & 1.4 \end{bmatrix} \begin{bmatrix} 15 \\ -5 \end{bmatrix} = \begin{bmatrix} 75 \\ -25 \end{bmatrix} = 5 \begin{bmatrix} 15 \\ -5 \end{bmatrix}.$$

It is common either to make a simple choice such as $\begin{bmatrix} 3 \\ -1 \end{bmatrix}$, or to choose an eigenvector with length

1, which in this case is either $\begin{bmatrix} \frac{3}{\sqrt{10}} \\ -\frac{1}{\sqrt{10}} \end{bmatrix}$ or $\begin{bmatrix} -\frac{3}{\sqrt{10}} \\ \frac{1}{\sqrt{10}} \end{bmatrix}$. (This is known as *normalizing*.)

For the eigenvalue 1, a simple eigenvector is $\begin{bmatrix} 1 \\ 3 \end{bmatrix}$ and a normalized eigenvector is $\begin{bmatrix} \frac{1}{\sqrt{10}} \\ \frac{3}{\sqrt{10}} \end{bmatrix}$.

Note that

- (i) eigenvectors are *orthogonal*, that is that $\mathbf{x}_i^T \mathbf{x}_j = 0$ for $i \neq j$, and
- (ii) when the eigenvectors have been chosen to be of length 1, then $\mathbf{x}_i^T \mathbf{x}_i = 1$ for all i .

For instance, in Example II.D.6, the two eigenvectors were perpendicular (the layman's

interpretation of orthogonal), and indeed $[1, 1] \begin{bmatrix} 1 \\ -1 \end{bmatrix} = 0$.

II.D.5.3 Determinants

For a 2×2 square matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ the quantity $ad - bc$ is called the *determinant* of the matrix.

Thus the determinant of a 2×2 matrix is defined by

$$\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc.$$

The determinant of a 3×3 matrix is defined by

$$\det\left(\begin{bmatrix} a & b & c \\ p & q & r \\ x & y & z \end{bmatrix}\right) = a \begin{vmatrix} q & r \\ y & z \end{vmatrix} - b \begin{vmatrix} p & r \\ x & z \end{vmatrix} + c \begin{vmatrix} p & q \\ x & y \end{vmatrix}.$$

This can be thought of as:

- $a \times$ the determinant given by crossing out a 's row and column
- $- b \times$ the determinant given by crossing out b 's row and column
- $+ c \times$ the determinant given by crossing out c 's row and column.

Higher -order determinants can be defined in a similar, recursive way, expanding by the first row (in fact by any row or column) with the signs alternating.

So, for the matrices which we have been using,

$$\begin{vmatrix} 0 & -1 \\ 1 & 0 \end{vmatrix} = 0 \times 0 - (-1 \times 1) = 1$$

and

$$\begin{vmatrix} 2 & 0 \\ 0 & 1 \end{vmatrix} = 2 \times 1 - 0 \times 0 = 2,$$

and for the 3×3 matrix $\begin{bmatrix} 2 & 3 & 1 \\ 4 & 5 & 1 \\ 6 & -2 & 7 \end{bmatrix}$,

$$\begin{vmatrix} 2 & 1 & 3 \\ 4 & 5 & 1 \\ 1 & -2 & 7 \end{vmatrix} = 2 \begin{vmatrix} 5 & 1 \\ -2 & 7 \end{vmatrix} - \begin{vmatrix} 4 & 1 \\ 1 & 7 \end{vmatrix} + 3 \begin{vmatrix} 4 & 5 \\ 1 & -2 \end{vmatrix} = 2(35 - (-2)) - (28 - 1) + 3(-8 - 5) = 74 - 27 - 39 = 8,$$

The determinant of a 2×2 matrix gives the 'area scale factor' of the transformation represented by the matrix, i.e. for any shape the area of its image under the transformation divided by the area of the shape. Similarly the determinant of a 3×3 matrix gives a 'volume scale factor'.

II.D.5.4 The Characteristic Equation

The process of finding eigenvalues can be shortened by using the *characteristic equation*. This is derived as follows:

$$\mathbf{Ax} = \lambda\mathbf{x} \Rightarrow (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}, \text{ where } \mathbf{I} \text{ is the } n \times n \text{ identity matrix.}$$

Now this can only have a non-zero solution for \mathbf{x} if $|\mathbf{A} - \lambda\mathbf{I}| = 0$, remembering that $|\mathbf{A} - \lambda\mathbf{I}|$ denotes the *determinant* of $\mathbf{A} - \lambda\mathbf{I}$ (see above). Thus

$$|\mathbf{A} - \lambda\mathbf{I}| = 0 \text{ is the } \textit{characteristic equation} \text{ of } \mathbf{A}.$$

Example II.D.8:

$$\mathbf{A} = \begin{bmatrix} 4.6 & -1.2 \\ -1.2 & 1.4 \end{bmatrix}, \quad \mathbf{A} - \lambda\mathbf{I} = \begin{bmatrix} 4.6 - \lambda & -1.2 \\ -1.2 & 1.4 - \lambda \end{bmatrix},$$

$$|\mathbf{A} - \lambda\mathbf{I}| = (4.6 - \lambda)(1.4 - \lambda) - 1.44 = \lambda^2 - 6\lambda + 5 = (\lambda - 5)(\lambda - 1).$$

So, as found before, the eigenvalues are $\lambda = 5$ and $\lambda = 1$.

II.D.5.4.1 Testing for Positive Semi-definiteness

Portfolio market risk assessment often requires the use of a covariance matrix. Sometimes this matrix is estimated using historical data (see Chapter III.A.3) and at other times it can be ‘made up’, for instance by changing risk factor correlations (see Chapter III.A.4). In the latter case – and even in the former case if an inexperienced user has done the estimation – there is a chance that the matrix will not be positive semi-definite (see Section II.D.3.2). But in that case:

- (a) some portfolio could have negative variance, and
- (b) there will be no Cholesky matrix to use in simulations

Hence it is often necessary to perform a simple test, on the matrix that is being used to assess the portfolio risk, to ensure that it is positive semi-definite. If an $n \times n$ matrix can be regarded as n independent scalings then it has a ‘full set’ of n eigenvectors/eigenvalues. Thus we have:

A matrix is positive definite if and only if it has a full set of positive eigenvalues.

A matrix is positive semi-definite if and only if it has a full set of non-negative eigenvalues.

What distinguishes the two cases is that positive semi-definite matrices can have zero eigenvalues.

Hence a simple test of positive (semi)definiteness is to calculate the eigenvalues of a matrix and make sure that they are all positive (or, at least, that none are negative).⁷

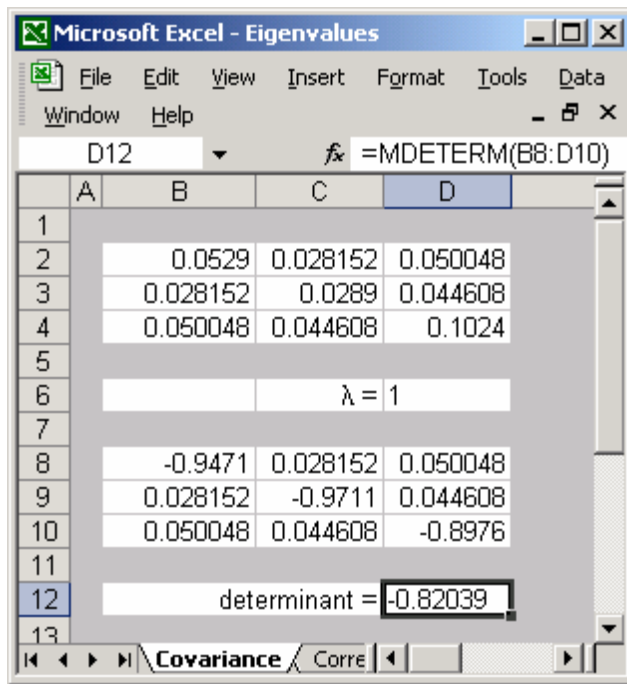
⁷ Note that in large matrices some eigenvalues that are really zero may in fact be calculated as very small negative numbers, due to rounding errors.

II.D.5.4.2 Using the characteristic equation to find the eigenvalues of a covariance matrix

Consider the covariance matrix $\mathbf{V} = \begin{bmatrix} 0.052900 & 0.028152 & 0.050048 \\ 0.028152 & 0.028900 & 0.044608 \\ 0.050048 & 0.044608 & 0.102400 \end{bmatrix}$.

The spreadsheet in Figure II.D.13 was set up to find the eigenvalues.

Figure II.D.13: Finding the eigenvalues of a matrix in Excel



Cells B2 to D4 contain the covariance matrix.

Cells B8 to D10 contain the same matrix with the number in D6 subtracted from each of B2, C3 and D4.

Cell C12 contains the determinant of this second matrix.

We then need to iterate, changing D6 until D12 is zero. Unfortunately ‘Goal Seek’ is not accurate enough for this task. However, ‘Solver’ can be used.

Solver's accuracy can be adjusted through its options. Furthermore, once one eigenvalue has been found it can be excluded from a subsequent search by adding in constraints (see Figure II.D.14 below).

II.D.5.4.3 Eigenvalues and eigenvectors of covariance and correlation matrices

In the 2×2 case it is very easy to see what the eigenvalues of a correlation matrix are. Indeed, following Example II.D.6 above we know that for a correlation matrix

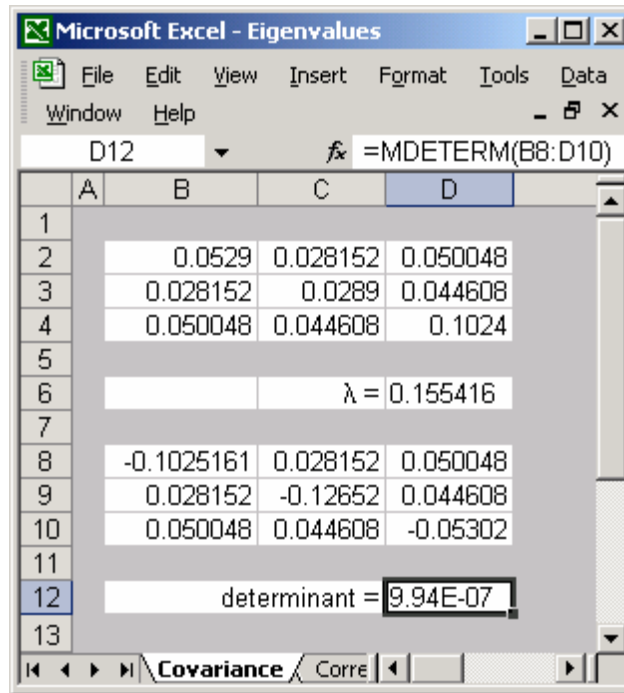
$$\begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = (1+r) \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

and

$$\begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = (1-r) \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

so that the eigenvalues are $(1 + r)$ and $(1 - r)$ with eigenvectors $(1, 1)^T$ and $(1, -1)^T$ respectively.

Figure II.D.14: Finding further eigenvalues when at least one is known



As shown in equation (II.D.1.11), the covariance and correlation matrices \mathbf{V} and \mathbf{C} are related by

$$\mathbf{V} = \mathbf{D}\mathbf{C}\mathbf{D}$$

where \mathbf{D} is the diagonal matrix of standard deviations. However, there is no simple relationship between their eigenvalues, as the following example shows.

Example II.D.9

A computer algebra package gave the eigenvalues of the covariance matrix \mathbf{V} above as 0.1554662298, 0.02182472434 and 0.006909045672.

The corresponding correlation matrix is $\begin{bmatrix} 1 & 0.72 & 0.68 \\ 0.72 & 1 & 0.82 \\ 0.68 & 0.82 & 1 \end{bmatrix}$.

The eigenvalues of this matrix are not related to the eigenvalues of the covariance matrix. They are 2.481601590, 0.3419189270 and 0.1764794869.

II.D.5.5 Principal Components

Consider a covariance matrix \mathbf{V} , which we know should always be positive definite. Thus \mathbf{V} should always have a full set of (positive) eigenvalues. In this case, the eigenvectors are known as

components. A component is essentially a direction (a linear combination of the variables). Let the eigenvectors be $\mathbf{x}_1, \mathbf{x}_2, \dots$, with eigenvalues $\lambda_1, \lambda_2, \dots$, etc., choosing the eigenvectors each to have length 1.

Suppose that \mathbf{w} is the vector of the weightings of individual assets in a portfolio. A theorem of linear algebra says that we can express any vector as a weighted sum of the eigenvectors and any matrix. So, by this theorem, we can write

$$\mathbf{w} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \dots + a_n\mathbf{x}_n.$$

Now the variance of the returns to the portfolio is given by $\mathbf{w}^T\mathbf{V}\mathbf{w}$ (see, for example, Section II.D.3.1). Thus the portfolio variance is

$$\begin{aligned}\mathbf{w}^T\mathbf{V}\mathbf{w} &= \mathbf{w}^T\mathbf{V}(a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \dots + a_n\mathbf{x}_n) \\ &= \mathbf{w}^T(a_1\lambda_1\mathbf{x}_1 + a_2\lambda_2\mathbf{x}_2 + \dots + a_n\lambda_n\mathbf{x}_n) \\ &= \lambda_1a_1^2 + \lambda_2a_2^2 + \dots + \lambda_na_n^2.\end{aligned}$$

Thus the variance is also split into components, and often most of the variance is accounted for by the first two or three *principal* components. This serves merely as a mathematical introduction to principal component analysis (PCA). In fact we shall see that PCA has many financial applications, to interest-rate modelling, market risk assessment, scenario analysis and much more.

References

Markowitz, H (1952) Portfolio selection, *Journal of Finance*, 7(1), pp. 77–91.

Watsham, T J, and Parramore, K (1997) *Quantitative Methods in Finance*. London: International Thomson Business Press.

II.E Probability Theory in Finance

Keith Parramore and Terry Watsham¹

Probability theory concerns the mathematical models that attempt to capture the behaviour of observed phenomena. In finance the ‘observed phenomena’ are usually associated with asset prices or interest rates. They can be the changes in these over a period of time, or just the changes that produce losses or percentage returns. Thus we use probability to estimate the certainty (or uncertainty) of the prices or rates behaving in a certain way. For example, what is the chance that the FTSE 100 will fall 2% over the next day, or what is the probability that the USD/CHF exchange rate will reach 2.00 by next December?

Probability is measured on a range from zero (complete certainty that an event will not happen) to one (absolute certainty that an event will happen). If the probability is either one or zero the event happens (or does not happen) with certainty and we say the event is *deterministic*. If the probability is in between zero and one, we say that the event is *probabilistic* or *stochastic*.

By the end of this Chapter you will:

- understand the concept of probability and the different approaches to defining and measuring it;
- be able to apply the rules of probability;
- understand the mathematics relating to univariate distributions, that is, probability functions for discrete random variables and probability density functions for continuous random variables, and the algebra of random variables;
- be able to compute the expectation and variance of a random variable;
- understand the mathematics relating to jointly distributed random variables;
- be able to compute a covariance, a correlation and the expected value and variance of a linear combination of random variables;
- understand and be able to use the binomial and Poisson (discrete) probability distributions;
- understand and be able to use the uniform, normal and lognormal (continuous) probability distributions;
- understand the form and use of the t distribution and the bivariate normal distribution.

¹ University of Brighton, UK.

II.E.1 Definitions and Rules

II.E.1.1 Definitions

There are two approaches to probability: classical and Bayesian.

II.E.1.1.1 The classical approach

The classical approach applies when probabilities are determined by an ‘experiment’ in which a range of possible ‘uncertain outcomes’ is known. The archetypal example is the tossing of a fair six-sided die, where each side is equally likely to come up.

- Throwing the die once is the *experiment*.
- An *outcome* is the result of one experiment: in this example, the outcome is the face showing uppermost. If the experiment is yet to be performed we refer to ‘possible outcomes’ or ‘possibilities’ for short. If the experiment has been performed, we refer to ‘realised outcomes’ or ‘realisations’ for short.
- The *sample space* is the set of possible outcomes: that is, {1, 2, 3, 4, 5, 6} in this example
- An *event* is a specific outcome or combination of outcomes. For example, an event associated with this example could be ‘the die shows an even number’.

The probability of an event occurring is defined as

$$P(A) = \frac{\text{No. of equally likely outcomes associated with the event}}{\text{Total number of equally likely outcomes}} \quad (\text{II.E.1})$$

In this example, as the die is assumed to be fair the probability of each face showing must be 1/6.

Thus if A is the event *the die shows an even number*, then

- the total number of equally likely outcomes is 6, as the die has 6 different faces numbered 1, 2, ..., 6;
- the number of equally likely outcomes associated with the event is 3, as the even numbers are 2, 4 and 6,

and

$$P(A) = 3/6 = 0.5.$$

The probability of A not occurring is $P(\text{not } A) = 1 - P(A)$. For instance, the probability of the die showing an odd number is also 0.5.

In the above example the symmetry of the die determined the probability of each outcome. But in finance, as in many other fields, we cannot rely on the existence of symmetry to determine probabilities. Where there is no underlying symmetry we may have to repeat an experiment many times to determine an experimental probability for each of a number of possible outcomes. The

range of possibilities for the return of a financial asset is virtually unlimited, thus financial analysts would have to observe many movements in asset prices in order to determine a probability to associate with future price changes of a given magnitude. In such situations the probability of a given outcome Z , $P(Z)$, is calculated as the ratio of the number of times that Z occurs to the number of times the experiment is conducted:

$$P(Z) = \frac{\text{No. of } Z \text{ occurrences}}{\text{No. of experiments}}. \quad (\text{II.E.2})$$

The reader may recognise this as the *relative frequency* of Z . This approach involves the analysis of historical data to estimate the probabilities that can be assigned to events. To illustrate this, assume that of 100 absolute price movements observed in the past, 5 movements were 0.5 cents each, 15 were 1 cent each, 20 were 1.5 cents, 30 were 2 cents, 20 were 2.5 cents and 10 were 3 cents. Then the probability that on a randomly chosen day in the period the price change was 1 cent is $15/100$ or 0.15. We can write this as $P(\text{change} = 1 \text{ cent}) = 0.15$. Adding up all the probabilities for all the price changes, $0.05 + 0.15 + 0.20 + 0.30 + 0.20 + 0.10 = 1$.

In general, the probability of a single outcome lies somewhere between zero and one, and the sum of the probabilities of a given set of outcomes must be one.

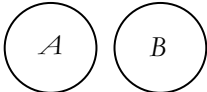
II.E.1.1.2 The Bayesian approach

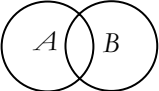
There is a second approach to probability, known as the ‘Bayesian’ approach after Thomas Bayes, a seventeenth-century English Presbyterian minister, who laid the foundations for all modern statistics. Probability under this approach is determined, at least in part, by a *subjective* ‘belief’ about the probability with which an event will occur. Subjective probability is applied to many problems, because empirical observations form only part of the information for future probability estimates. For example, subjective probability may be incorporated into the forecasting of company profits by investment analysts. As more evidence becomes available these probability estimates may be adjusted by applying a result known as *Bayes’s Theorem*.

II.E.1.2 Rules for Probability

II.E.1.2.1 (A or B) and (A and B)

If A and B are events occurring with probabilities $P(A)$ and $P(B)$ respectively, then we may wish to know the probability of either event A or event B occurring. The probability of the ‘compound’ event A or B is written $P(A \vee B)$. It is helpful here to consider a picture of our events.

They might look like this: 

or they might look like this: 

In the first case there is no possibility of the events occurring simultaneously, in the second case there is such a possibility. The probability of both events occurring simultaneously is the probability of A and B . It is written as $P(A \wedge B)$. In the first case $P(A \vee B) = P(A) + P(B)$. In the second case adding the two probabilities would involve counting the conjunction ($A \wedge B$) twice. Therefore the correct formula is:

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B) \quad (\text{II.E.3})$$

Of course, this formula applies to both cases, since in the first case $P(A \wedge B) = 0$.

Example II.E.1:

Suppose that we have the following historical data on the returns to two investments:

- Number of years in which both recorded a gain = 12
- Number of years in which investment 1 gained and investment 2 lost = 3
- Number of years in which investment 2 gained and investment 1 lost = 2
- Number of years in which both investments recorded losses = 3

Let A be the event ‘investment 1 records a gain’. Let B be the event ‘investment 2 records a gain’.

Then, using empirical probabilities,

$$P(A) = 0.75, P(B) = 0.7, P(A \wedge B) = 0.6$$

and

$$P(A \vee B) = 0.75 + 0.7 - 0.6 = 0.85.$$

II.E.1.2.2 Conditional Probability

Although $P(A \wedge B)$ appears in the formula for $P(A \vee B)$, we will also need to be able to compute it in terms of underlying probabilities. To do this we need to introduce the concept of *conditional probability*.

Suppose that in Example II.E.1 we are given the information that investment 1 has returned a gain. Given that, what is the probability that investment 2 has returned a gain?

The notation for this is $P(B \mid A)$, and is read as ‘the probability of B given A ’.

Effectively the information that A has returned a gain restricts us to considering a subset of the possibilities – in this case represented by those 15 years in which A did show a gain. Out of those 15 years B also showed a gain in 12, so $P(B \mid A) = 12/15 = 0.8$.

We can now give the formula for $P(A \wedge B)$. It is:

$$P(A \wedge B) = P(B | A) \times P(A) \quad (\text{II.E.4})$$

Of course, it is also true that $P(A \wedge B) = P(A | B) \times P(B)$. Which of the two formulae is used depends on the circumstances, and on the labelling of events!

II.E.1.2.3 Independent Events

It may happen that $P(B | A) = P(B)$ (or $P(A | B) = P(A)$). In such circumstances we say that A and B are *independent*. If A and B are independent then:

$$P(A \wedge B) = P(A) \times P(B). \quad (\text{II.E.5})$$

In Example II.E.1, $P(B | A) = 0.8$ and $P(A) = 0.75$. So $P(A \wedge B) = 0.8 \times 0.75 = 0.6$, and A and B are not independent.

II.E.2 Probability Distributions

Where an experiment is repeated many times it is convenient not just to consider specific possibilities and their associated probabilities, but all possibilities and probabilities. Taken together these are known as the *probability distribution* of a *random variable*.

II.E.2.1 Random Variables

In Chapter II.B, on statistics, we implicitly assumed that we knew about the (historical) behaviour of the variable being analysed, and that we were summarising that historical behaviour by computing appropriate statistics. We were describing the past behaviour. In this chapter we will model the uncertain, that is, ‘random’ behaviour of a variable. In effect, we are looking into the future and the variable of interest is a *random variable*. A random variable is a variable that behaves in an uncertain manner.

We will often use the historical behaviour of a variable to *calibrate* our model of the random (uncertain future) behaviour of that variable. As this behaviour is uncertain we can only assign probabilities to the possible values which the variable can take. Our models of random behaviour are *probability distributions*. In summary, the random variable is defined by its possible outcomes and by the probabilities of those outcomes.

We have *discrete* probability distributions for explaining the behaviour of discrete random variables: these are random variables that can take only certain discrete values. For example, the throw of a single die can only give a discrete number of integer outcomes, from one to six. We

have *continuous* probability distributions for explaining the behaviour of continuous random variables. Here the variable can take on any value within a specified range, including non-integers. We begin with discrete random variables.

II.E.2.1.1 Discrete Random Variables

Discrete random variables are those that have only a discrete set of possible outcomes. They relate to situations involving counting rather than measurement. Consider again the situation when a six-sided die is thrown. Each of the possible outcomes has a probability associated with it. If the die is unbiased, each of those six probabilities is 1/6. This process is modelled mathematically as a discrete random variable.

We could call the random variable Z and define it thus:

Possibility(z)	1	2	3	4	5	6
$P(Z = z)$	1/6	1/6	1/6	1/6	1/6	1/6

The possibilities, together with their associated probabilities, constitute the *probability distribution*.

Note that the probabilities, irrespective of how many possibilities there are, must always sum to one, i.e.

$$\sum_z P(Z = z) = 1 . \tag{II.E.6}$$

(Recall from Chapter II.A that Σ indicates a summation, and that \sum_z means the sum over all possible values of z .)

Examples of discrete distributions include the binomial distribution (Section II.E.4.1) and the Poisson distribution (Section II.E.4.2).

II.E.2.1.2 Continuous Random Variables

Continuous random variables are those which relate to measuring rather than to counting. Distance, speed, time and asset returns are examples. The unit of measurement can be increased or decreased by infinitesimally small amounts. The number of possible outcomes for this particular random variable is uncountably infinite. Under these circumstances it makes no sense to consider the possibility of the random variable taking a specified value. It only makes sense to consider the possibility of the continuous random variable taking values *between two limits*.

For example, the return from holding a security is, for our purposes, virtually a continuous random variable. Thus it is not appropriate to consider the possibility of the return taking the

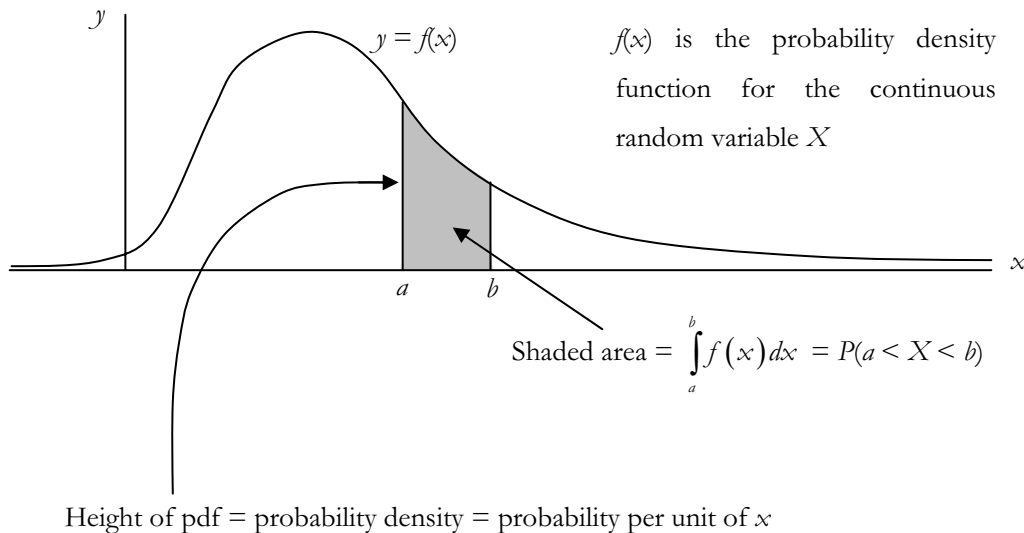
value of a specific value such as (say) 3.812403%. It only makes sense to consider the possibility of the return taking values between two limits, for example the possibility that the return will be between 3.80% and 3.82%.

We cannot define a continuous random variable by listing its possibilities and their associated probabilities. Instead it is defined by its *probability density function* (pdf). For a continuous random variable, X , its probability density function, $f(x)$, has the two properties that

$$f(x) \geq 0 \text{ for all } x (-\infty < x < \infty) \text{ and } \int_{-\infty}^{\infty} f(x) dx = 1 \quad (\text{II.E.7})$$

A specific probability, the probability that X is between two given values, is then defined by the corresponding *area under the pdf* (Figure. II.E.1).

Figure.II.E.1: Probability of a continuous random variable



Note that the height of a probability density function at a point given by $x = a$ does *not* represent $P(X = a)$. For a continuous random variable it is not meaningful to consider the probability of that random variable taking a specific value, only of it lying within a range of values.

II.E.2.2 Probability Density Functions and Histograms

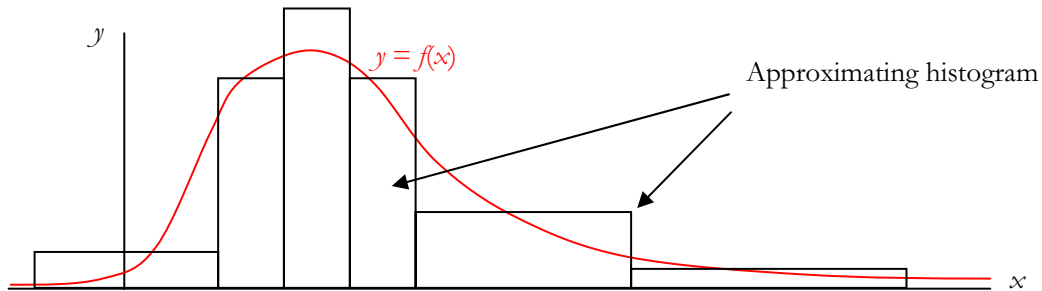
For f to be a pdf,

$$f(x) \text{ must be non-negative for all possible values of } x, \text{ and } \int_{-\infty}^{\infty} f(x) dx = 1.$$

If data are collected from realisations of a continuous random variable then observations are collected into ‘bins’. Relative frequencies are computed and rectangles are constructed for each bin so that rectangle area equals relative frequency. The resulting graph is called a *histogram*. It

represents a graphical estimate of the unknown pdf. Looking at Figure II.E.2, we see that the finer the divisions of the horizontal axis, the more closely will the tops of the rectangles match the curve.

Figure II.E.2: The histogram and the pdf



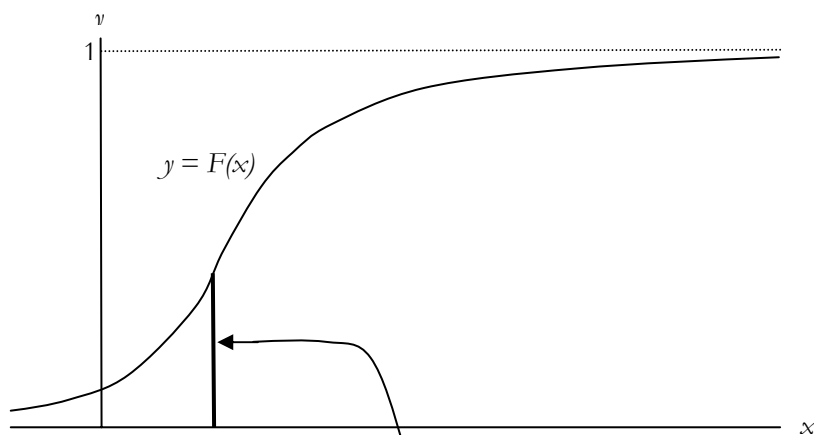
II.E.2.3 The Cumulative Distribution Function

In layman's terms, the *cumulative distribution function* (cdf) gives the 'area so far'. Mathematically, the cdf is the function $F(x)$ given by

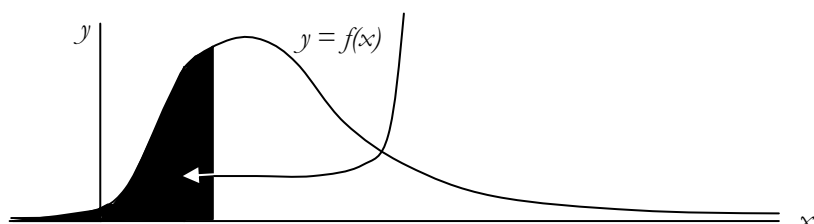
$$F(x) = \int_{-\infty}^x f(u) du .$$

The cdf is constructed from the pdf as shown in Figure II.E.3.

Figure II.E.3: The cumulative distribution function



The height of the cumulative distribution function, $F(x)$ is the 'area so far' underneath the probability density function $f(x)$.



Note 1: Recall from Chapter II.C that $\int \dots du$ stands for integration with respect to u , and that this is the continuous equivalent of summation.

Note 2: In this expression u is a ‘dummy’ variable. It is x that is the variable. The u is a notational device for covering all of the possibilities up to x .

You will find that this concept is particularly important when dealing with the *normal distribution* (see Section II.E.4.4). It appears in the Black–Scholes equation (see Section I.A.8.7) where the N in ‘ $N(d_1)$ ’ and ‘ $N(d_2)$ ’ is the cumulative (standard) normal distribution.

II.E.2.4 The Algebra of Random Variables

II.E.2.4.1 Scalar Multiplication of a Random Variable

When multiplying a discrete random variable by a fixed number, the probabilities remain unchanged but the possibilities are multiplied by that fixed number. For example, if X is a discrete random variable, then $2X$ is defined by the same probability distribution as X , except that the possibilities are all doubled (the probabilities remain the same).

Thus if X is defined by:

Possibility (r)	0	1	2
Probability that $X = r$	1/4	1/2	1/4

then $2X$ is defined by:

Possibility (r)	0	2	4
Probability that $X = r$	1/4	1/2	1/4

II.E.2.4.2 Adding Two Independent Random Variables

Now we can consider the sum of two discrete random variables, X and Y . Suppose that

X is defined by:

Possibility (r)	0	1	2
Probability that $X = r$	1/4	1/2	1/4

Y is defined by:

Possibility (r)	4	5
Probability that $Y = r$	1/2	1/2

and that X and Y are *independent* (see Section II.E.1.2.3).

By computing the possibilities and their associated probabilities, we can show that $X + Y$ is given by:

Possibility (r)	4	5	6	7
Probability that $X + Y = r$	1/8	3/8	3/8	1/8

First the possibilities:

$$0 + 4 = 4, \quad 0 + 5 = 5 \quad 1 + 4 = 5, \quad 1 + 5 = 6, \quad 2 + 4 = 6, \quad 2 + 5 = 7$$

Thus the possibilities are: 4 (one way), 5 (two ways), 6 (two ways) and 7 (one way).

One possibility is 4, from $0 + 4$. This has a probability of $1/4 \times 1/2 = 1/8$ (using equation (II.E.5)).

Next is the possibility of 5, achieved in two ways, $0 + 5$ with probability $1/4 \times 1/2 = 1/8$, or $1 + 4$ with probability $1/2 \times 1/2 = 1/4$. Adding the probabilities for each possibility (see equation (II.E.3)) gives a probability of $1/8 + 1/4 = 3/8$.

For each of the remaining possibilities the corresponding probabilities can be computed in a similar way.

Note: For random variables, $2X$ is *not the same* as $X + X$! (Compare $2X$ and $X + X$ in the examples above.)

II.E.2.5 The Expected Value of a Discrete Random Variable

The expected value (or expectation) of a discrete random variable is defined by

$$E(X) = \sum_r (r \times P(X = r)). \quad (\text{II.E.8})$$

E is known as the *expectation operator*. For example, our random variable X defined above (Section II.E.2.4) has an expected value of

$$E(X) = 0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1.$$

The expected value of a random variable is the *probabilistically weighted mean* of the various possible values which can be taken by our random variable. It should be noted that the expected value need not be a member of the set of possibilities. For example, consider the six-sided die used earlier. The expected value of a single outcome is

$$\left(1 \times \frac{1}{6}\right) + \left(2 \times \frac{1}{6}\right) + \left(3 \times \frac{1}{6}\right) + \left(4 \times \frac{1}{6}\right) + \left(5 \times \frac{1}{6}\right) + \left(6 \times \frac{1}{6}\right) = 3.5,$$

but 3.5 is not a possible outcome when the die is thrown!

II.E.2.6 The Variance of a Discrete Random Variable

Variance was defined in Chapter II.B and Section I.A.2.2. Corresponding to this, the *variance operator* is defined by

$$\text{Var}(X) = E(X - E(X))^2. \quad (\text{II.E.9})$$

Note that this definition applies to both *discrete* and *continuous* random variables, X .

For an example of the variance of a discrete random variable, consider the expected return from a particular asset. Suppose that an investment manager considers that over a given future period one of three economic scenarios, or states of the world, will apply: high growth, no growth or a recession. The probabilities of each scenario occurring are believed to be 0.25, 0.5 and 0.25, respectively (readers will recognize this as an example of subjective probability). The investment manager expects a particular asset to earn 20% if high growth prevails, 10% if no growth prevails and -4% if there is a recession. As the expected value of a random variable is the value of each expected outcome multiplied by the respective probability, the expected return of the asset in question would be

$$E(R) = (0.20 \times 0.25) + (0.10 \times 0.50) - (0.04 \times 0.25) = 0.09 = 9\%.$$

The variance of a random variable is *the sum of the products of the squared deviations from the expected outcome, multiplied by their respective probabilities*. In this example we have:

$$\begin{aligned} \text{Var}(R) &= (0.20 - 0.09)^2 \times 0.25 + (0.10 - 0.09)^2 \times 0.50 + (-0.04 - 0.09)^2 \times 0.25 \\ &= 0.003025 + 0.00005 + 0.004225 = 0.0073 \end{aligned}$$

This outcome is in percentages *squared*, which is not intuitively very appealing, so the square root of the variance, the *standard deviation*, is usually reported. This will always be in the same units of measurement as the returns. In this example, the square root of 0.0073 is 0.085 or 8.5%.

Thus we see that if asset returns are a random variable, the expected return is the probabilistically weighted mean of the expectation of returns, and the risk, as measured by the variability of those expectations, is described by the variance or standard deviation. Note that there is an alternative expression for the variance operator:

$$\text{Var}(X) = E(X^2) - (E(X))^2. \quad (\text{II.E.10})$$

In the example above we have:

$$\begin{aligned} \text{Var}(X) &= [0.20^2 \times 0.25 + 0.10^2 \times 0.50 + (-0.04)^2 \times 0.25] - 0.09^2 \\ &= [0.01 + 0.005 + 0.0004] - 0.0081 \\ &= 0.0154 - 0.0081 = 0.0073, \text{ as given above.} \end{aligned}$$

II.E.2.7 The Algebra of Continuous Random Variables

The expected value of a continuous random variable X is given as

$$E(X) = \mu_X = \int_{-\infty}^{+\infty} xf(x) dx, \text{ where } f \text{ is the pdf of } X. \quad (\text{II.E.11})$$

If X is limited to a range of possible values, say between possibilities x_1 and x_2 , the expected value simplifies to

$$E(X) = \mu_X = \int_{x_1}^{x_2} xf(x) dx. \quad (\text{II.E.12})$$

The variance and standard deviation of a continuous random variable X are

$$\text{Var}(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu_X^2 \quad (\text{II.E.13})$$

and

$$\sigma_X = \sqrt{\text{Var}(X)}. \quad (\text{II.E.14})$$

Note that (II.E.9) and (II.E.10) still hold for the continuous random variable X .

Whilst the distinction between the discrete and the continuous is vital in building and understanding theoretical models, we can sometimes use a continuous random variable to model a discrete situation (and vice versa). For instance, consider the price of a particular share on the stock exchange at midday on the next trading day. Clearly there *are* only a discrete set of possibilities (shares are priced in pounds sterling, pence and occasionally fractions of pence). Nevertheless it may well be that a continuous random variable may provide the best model of the behaviour of that share price.

II.E.3 Joint Distributions

Often we need to consider the joint behaviour of two or more random variables. For instance, in the capital asset pricing model (see Chapter I.A.4) we consider the return on an equity and the return on a market index. For pricing convertible bonds or other hybrid instruments (see Chapter I.B.8) we are interested in the joint behaviour of the issuer's equity, the zero coupon rate and the credit spread. For simulating correlated returns in Monte Carlo value-at-risk (see Chapter III.A.2) we require knowledge of the joint behaviour of the portfolio's risk factor returns: these can be interest rates of different maturities (for a loan portfolio, for instance); broad stock market indices and foreign exchange rate (for an international equity portfolio); commodity future returns and so forth. In such cases we are interested in their *joint probability distribution*.

II.E.3.1 Bivariate Random Variables

Consider two discrete random variables, X and Y . X can take values 1, 2 or 4 with probabilities 0.2, 0.5 and 0.3 respectively. Y can take values 20 or 30 with probabilities 0.5 and 0.5. Realisations occur simultaneously, producing a pair of values (x, y) . For instance, X could represent future percentage returns to an asset, the probabilities being determined subjectively, whereas Y could represent the possible future volatility of the asset, again with probabilities being determined subjectively.

The table shows the set of possible outcomes. The asterisks in the centre represent probabilities which define the *joint distribution* of X and Y . Taken together this may also be referred to as a *bivariate random variable*.

		X			
		1	2	4	
Y	20	***	***	***	0.5
	30	***	***	***	0.5
		0.2	0.5	0.3	

If the two random variables are *statistically independent*² then the realisations of X are determined independently of the realisations of Y , and vice versa. The probabilities in the cells are then the products of the probabilities in the margins, as in case 1 below.

Case 1:

		X			
		1	2	4	
Y	20	0.1	0.25	0.15	0.5
	30	0.1	0.25	0.15	0.5
		0.2	0.5	0.3	

e.g., $0.15 = 0.5 \times 0.3$

² We distinguish between *statistical independence*, which is to do with probabilities, and *physical independence*, which is to do with linking mechanisms.

If the probabilities in the top-left and bottom-right cells (shaded in the diagram below) are larger than in the independent case, then this indicates that there is some mechanism which is causing the variables to vary in unison. In this case there is said to be *positive correlation*, as in case 2 below.

Case 2:

		X			
		1	2	4	
Y	20	0.15	0.3	0.05	0.5
	30	0.05	0.2	0.25	0.5
		0.2	0.5	0.3	

Note that this table has just *two degrees of freedom*. This is because you have freedom (though not unlimited freedom) in choosing entries for, say, the two shaded cells. But once entries are determined for those two cells, then entries in all other cells are defined by virtue of the required row and column totals.

If the probabilities in the bottom-left and top-right cells (shaded in the diagram below) are larger than in the independent case, then this indicates that there is some mechanism which is causing the variables to vary in opposition. In this case there is said to be *negative correlation*, as in Case 3 below:

Case 3:

		X			
		1	2	4	
Y	20	0.05	0.2	0.25	0.5
	30	0.15	0.3	0.05	0.5
		0.2	0.5	0.3	

II.E.3.2 Covariance

We need to have a measure which distinguishes between the three cases in Section II.E.3.1. The first step towards such a measure is to compute the *covariance* (compare what follows with Section I.A.2.2.2 and Chapter II.B).

$$\text{Cov}(X,Y) = E\left[(X - E(X)) \times (Y - E(Y))\right] = E(XY) - E(X)E(Y). \quad (\text{II.E.15})$$

Again, this definition applies to both discrete and continuous random variables.

In the examples above (cases 1–3), we have

$$E(X) = (1 \times 0.2) + (2 \times 0.5) + (4 \times 0.3) = 2.4 \quad \text{and} \quad E(Y) = (20 \times 0.5) + (30 \times 0.5) = 25.$$

This is the same in all cases, but $E(XY)$ differs according to the assumption about independence/positive correlation/negative correlation:

Case 1: Independence:

$$E(XY) = (1 \times 20 \times 0.1) + (2 \times 20 \times 0.25) + (4 \times 20 \times 0.15) + (1 \times 30 \times 0.1) + (2 \times 30 \times 0.25) \\ + (4 \times 30 \times 0.15) = 60.$$

So

$$\text{Cov}(X, Y) = 60 - 2.4 \times 25 = 0.$$

The covariance is zero if and only if the random variables are statistically independent, that is, if and only if there is no correlation between them.

Case 2: Positive covariance:

$$E(XY) = (1 \times 20 \times 0.15) + (2 \times 20 \times 0.3) + (4 \times 20 \times 0.05) + (1 \times 30 \times 0.05) + (2 \times 30 \times 0.2) \\ + (4 \times 30 \times 0.25) = 62.5.$$

So

$$\text{Cov}(X, Y) = 62.5 - 2.4 \times 25 = 2.5.$$

The covariance is positive if and only if there is positive correlation between the variables.

Case 3: Negative covariance:

$$E(XY) = (1 \times 20 \times 0.05) + (2 \times 20 \times 0.2) + (4 \times 20 \times 0.25) + (1 \times 30 \times 0.15) + (2 \times 30 \times 0.3) \\ + (4 \times 30 \times 0.05) = 57.5.$$

So

$$\text{Cov}(X, Y) = 57.5 - 2.4 \times 25 = -2.5.$$

The covariance is negative if and only if there is negative correlation between the variables.

II.E.3.3 Correlation

Recall that the variance may not have an intuitive interpretation as it is measured in the *square* of the unit of the variable. The same applies to covariance. Furthermore, the numerical value of the covariance will depend upon the units of measure (e.g., whether it is basis points or percentage points). With variance we restore sense by taking the square root to give the standard deviation. This is readily interpretable, though it still depends on the units of measurement. In the case of covariance we aid interpretation *and* standardise with respect to units of measurement by dividing by the product of the standard deviations to produce the *correlation coefficient*, ρ . This is

always between -1 and $+1$, perfect correlation corresponding to $\rho = 1$, statistical independence corresponding to $\rho = 0$, and perfect negative correlation corresponding to $\rho = -1$. Thus

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \quad (\text{II.E.16})$$

(This also should be compared to Sections I.A.2.1.4 in which the corresponding symbols used are r and s for sample values, compared to ρ and σ for parameters.)

In portfolio optimisation (see Section II.D.2 and Section I.A.2.3) we often use equation (II.E.16) in the form:

$$\text{Cov}(X, Y) = \rho_{XY} \sigma_X \sigma_Y \quad (\text{II.E.17})$$

Thus in the three cases shown in Section II.E.3.2 the correlation coefficients are computed as follows.³ Firstly,

$$\text{Var}(X) = E(X^2) - (E(X))^2 = (1 \times 0.2) + (4 \times 0.5) + (16 \times 0.3) - 2.4^2 = 7 - 5.76 = 1.24$$

so

$$\sigma_X = \sqrt{1.24} \approx 1.114,$$

and

$$\text{Var}(Y) = E(Y^2) - (E(Y))^2 = (400 \times 0.5) + (900 \times 0.5) - 25^2 = 650 - 625 = 25$$

so

$$\sigma_Y = \sqrt{25} = 5.$$

$$\text{Case 1: } \rho_{XY} = \frac{0}{1.114 \times 5} = 0 = 0$$

$$\text{Case 2: } \rho_{XY} = \frac{25}{1.114 \times 5} \approx 0.449$$

$$\text{Case 3: } \rho_{XY} = -\frac{25}{1.114 \times 5} \approx -0.449$$

³ Note: For comments on the significance of given values of the correlation coefficient see Chapter II.B.

II.E.3.4 The Expected Value and Variance of a Linear Combination of Random Variables

A *linear combination* of two random variables X and Y is the random variable $aX + bY$, where a and b are fixed numbers. This extends easily here, and in what follows, to more than two variables.

The expectation of a linear combination of random variables is the same linear combination of the expectation of each random variable. Thus:

$$E(aX + bY) = aE(X) + bE(Y). \quad (\text{II.E.18})$$

The variance, however, is nonlinear. It is given by:

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + 2ab\text{Cov}(X, Y) + b^2\text{Var}(Y). \quad (\text{II.E.19})$$

The result in equation (II.E.19) is of fundamental importance in portfolio analysis (see Chapter II.D).

If two random variables X and Y are uncorrelated, so that their covariance is zero, then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y). \quad (\text{II.E.20})$$

II.E.4 Specific Probability Distributions

Different random variables follow different probability distributions. In finance there are several specific univariate distributions which have significant roles. In this section we describe two discrete distributions (the binomial and the Poisson) five continuous distributions (the uniform, normal, lognormal, and Student t distributions) and one joint distribution (the bivariate normal).⁴

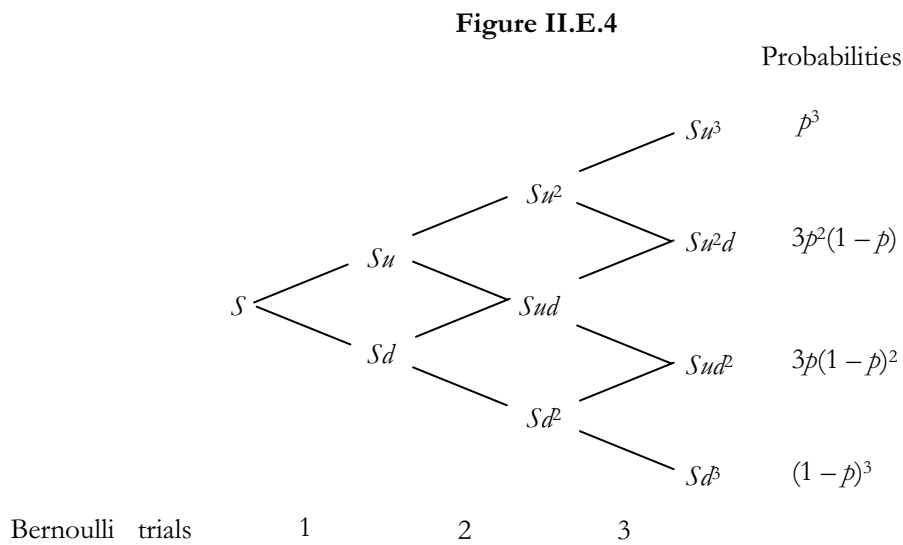
II.E.4.1 The Binomial Distribution

The *binomial distribution* is one of the most important discrete distributions in finance. It is used extensively in modelling the behaviour of asset prices. Consider a random variable which can take one of two possible values or outcomes. Each of these experiments is known as a (*Bernoulli*) *trial*. In each of a succession of such trials the probability of each of the two outcomes is constant, and each trial is independent of other trials. The two possible outcomes are sometimes referred to as ‘success’ and ‘failure’. In finance, a success would reflect an asset price rising, or positive returns, and a failure would reflect an asset price falling, or negative returns.

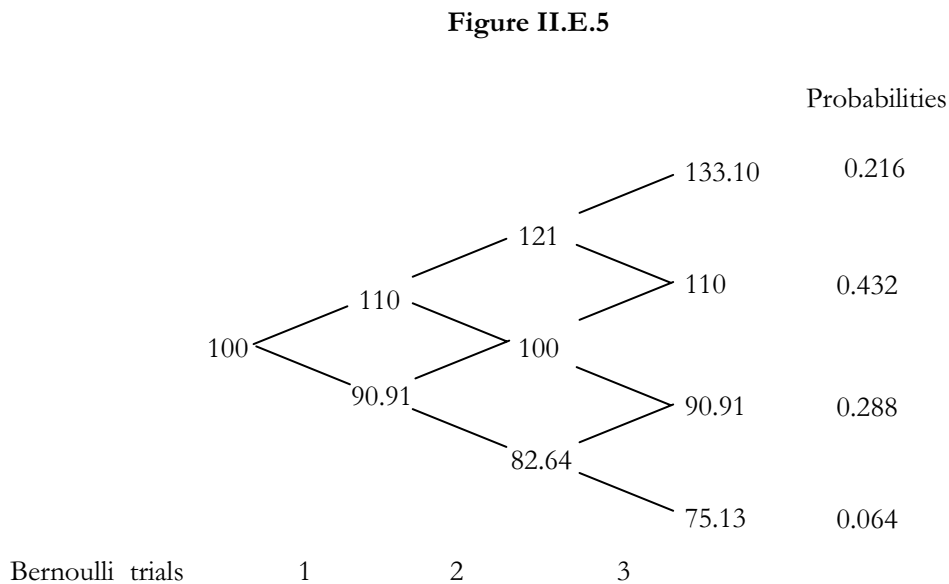
⁴ In constructing distributional models for observed phenomena we assume that realisations are independent and identically distributed. Whilst this is in many instances a reasonable starting point, the reader should bear in mind that it is not always the case for financial variables.

The binomial random variable, X , is the number of successes in n independent trials when the probability of success on each trial is a constant, p . We then say that $X \sim B(n, p)$. There are $n + 1$ possibilities for X (the values which it can take can take), namely $0, 1, 2, \dots, n$.

An application is illustrated in Figure II.E.4, where a security price, S , can move up by a factor of u ($u > 1$) or down by a factor of d ($0 < d < 1$). In this illustration there are three Bernoulli trials and therefore four outcomes. If r is used to count the number of upward movements (successes), then outcome Su^3 is the result of three successes, that is, $r = 3$. The outcome Su^2d is the result of two successes, so $r = 2$. The outcome Sud^2 is the result of one success so $r = 1$. Finally Sd^3 is the result of zero successes, so $r = 0$.



A particular realisation is shown in Figure II.E.5, with $S = 100$, $u = 1.1$, $d = 1/u \approx 0.09091$ and $p = 0.6$.



The binomial distribution gives the probabilities of each of these outcomes. The probability of achieving each outcome depends on:

1. the total number of ways of achieving that outcome.
2. the probability of achieving a success, p .

Thus in Figure II.E.5 the probability of the outcome 133.10 (Su^3) is the probability of three consecutive successes. Assume that the probability of a success is 0.6. Recalling that the trials are independent, the probability of three consecutive successes is given by $0.6 \times 0.6 \times 0.6 = 0.216$ (see equation (II.E.5)).

The probability of the outcome Su^2d is the probability of two successes and one failure. There are three ways to achieve this - two successes followed by a failure, a failure followed by two successes or a success followed by a failure which in turn is followed by another success. Each way has a probability of $0.6 \times 0.6 \times 0.4 = 0.144$, but as there are three ways, the total probability is $3 \times 0.144 = 0.432$ (see equation (II.E.3)).

The same reasoning can be used to show that the probability of the outcome Sud^2 is $3 \times 0.6 \times 0.4^2 = 0.288$ and that the probability of the outcome Sd^3 is $0.4^3 = 0.064$.

II.E.4.1.1 Calculating the ‘Number of Ways’

In the above example the number of ways of achieving Su^2d was easy to count. In general, we can calculate the number of ways of achieving r successes in a given number, n , of Bernoulli trials by using the following formula

$${}^n C_r = \frac{n!}{r!(n-r)!} \quad (\text{II.E.21})$$

The ! is *factorial notation*, and

$$n! = n(n-1)(n-2)(n-3)\dots 3 \times 2 \times 1.$$

For example $4! = 4 \times 3 \times 2 \times 1 = 24$.

Thus the number of ways of achieving Su^2d could have been computed as

$${}^3 C_2 = \frac{3!}{2!1!} = \frac{3 \times 2 \times 1}{(2 \times 1) \times 1} = 3.$$

Had we added on one more trial then we would have seen outcomes such as Su^2d^2 , and there are

$${}^4 C_2 = \frac{4!}{2!2!} = \frac{4 \times 3 \times 2 \times 1}{(2 \times 1) \times (2 \times 1)} = 6$$

ways of achieving this. Had we persisted to, say 20 trials, then we would have seen outcomes such as $Su^{14}d^6$. The number of ways of achieving this is

$${}^{20}C_{14} = \frac{20!}{14!6!} = \frac{20 \times 19 \times 18 \times \dots \times 3 \times 2 \times 1}{(14 \times 13 \times 12 \times \dots \times 3 \times 2 \times 1) \times (6 \times 5 \times 4 \times 3 \times 2 \times 1)} = \frac{20 \times 19 \times 18 \times 17 \times 16 \times 15}{6 \times 5 \times 4 \times 3 \times 2 \times 1} = 38,760$$

(note the cancellation of 14!).

Zero factorial, 0!, is defined to be 1, and this ensures that the formula works for $r = 0$ and for $r = n$, as well as for values in between.

II.E.4.1.2 Calculating the Probability of r Successes

38,760 is quite a large number of ways, but remember that there is a probability associated with each way. Since each involves 14 successes and 6 failures, each has the same probability. In the example the probability of a success was 0.6, so the probability of 14 successes and 6 failures is $0.6^{14} \times 0.4^6$. Thus the probability of achieving $Su^{14}d^6$ is $38,760 \times 0.6^{14} \times 0.4^6 = 0.124411699214785 \approx 0.1244$.

In general, we can calculate the probability of r successes by using the following formula

$$P(X = r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{(n-r)} \text{ for } 0 \leq r \leq n \quad (\text{II.E.22})$$

where p is the probability of success and $1 - p$ is the probability of failure.

Since the binomial process is the result of n trials there will be $n + 1$ outcomes, namely 0 successes, 1 success, 2, 3, ..., $n - 2$, $n - 1$ or n successes. In our example with three trials there are four possible outcomes, and the associated probabilities are tabulated below.

Possibility (value of r)	0	1	2	3
Outcome	Sd^3	Sud^2	Su^2d	Su^3
Probability	$\frac{3!}{0!3!} 0.6^0 0.4^3$ = 0.064	$\frac{3!}{1!2!} 0.6^1 0.4^2$ = 0.288	$\frac{3!}{2!1!} 0.6^2 0.4^1$ = 0.432	$\frac{3!}{3!0!} 0.6^3 0.4^0$ = 0.216

The probabilities can easily be produced in Excel, as shown in Figure II.E.6.

Figure II.E.6: [Calculating the probabilities for the binomial distribution in Excel](#)

	A	B	C	D
1	0	1	2	3
2	0.064	0.288	0.432	0.216
3				

II.E.4.1.3 Expectation and Variance

If $X \sim B(n, p)$ then:

$$E(X) = np; \quad \text{Var}(X) = np(1-p); \quad \sigma_X = \sqrt{np(1-p)} \quad (\text{II.E.23})$$

To confirm this in the $B(3, 0.6)$ example note that the expected number of successes is:

$$(0 \times 0.064) + (1 \times 0.288) + (2 \times 0.432) + (3 \times 0.216) = 0 + 0.288 + 0.864 + 0.648 = 1.8,$$

and that this is indeed equal to 3×0.6 .

Similarly, the variance of the number of successes is:

$$\begin{aligned} & (0^2 \times 0.064) + (1^2 \times 0.288) + (2^2 \times 0.432) + (3^2 \times 0.216) - 1.8^2 \\ & = 0 + 0.288 + 1.728 + 1.944 - 3.24 = 0.72, \end{aligned}$$

and that this is indeed equal to $3 \times 0.6 \times 0.4$.

Note that in the example it is the number of successes (upward movements), that is price *changes*, that is binomially distributed, and not the asset price itself.

II.E.4.2 The Poisson Distribution

Suppose that data entry errors occur in a back office randomly but at an unchanging average rate of 24 errors per (working) day (of 8 hours). How many errors might we see in a given hour?

This question is asking for a probability distribution. As a first step towards answering the question we might split our hour up into six 10-minute periods, with a probability of 0.5 of seeing an error in each period. This reflects the rate of 3 errors per hour or 24 errors per day. This is modelling the number of errors in the hour by a $\mathbf{B}(6, 0.5)$ distribution (see Section II.E.4.1).

Whilst this might be a good first approximation it cannot capture the essence of the process. For instance, it only allows for a maximum of six errors in our hour. There might be more.

An obvious improvement is to split the hour into 60 one-minute intervals, with a probability of 0.05 of an error occurring in any one interval, that is, a $B(60, 0.05)$ model.

Similarly we could improve again and again by constructing successive binomial models, $B(120, 0.025)$, $B(240, 0.0125)$, etc. The probabilities given by these models tend to limits, and these limits form a new probability distribution, the Poisson distribution.⁵

If X represents the number of occurrences of an event over a time period t , when occurrences are happening at an average rate of λ , then the probability distribution of X is given by:

$$P(X = r) = \frac{(\lambda t)^r e^{-\lambda t}}{r!} \quad (\text{II.E.24})$$

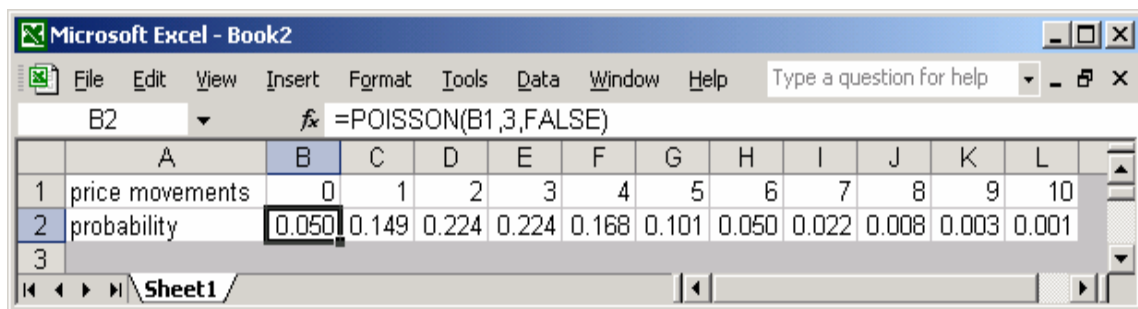
where $e = 2.7182\dots$ ⁶ The expected value, $E(X)$, is equal to λt (see equation (II.E.26) below), and we therefore have $\mu = \lambda t$. Thus equation (II.E.24) is often written in the form

$$P(X = r) = \frac{\mu^r e^{-\mu}}{r!} \quad (\text{II.E.25})$$

II.E.4.2.1 Illustrations

We use Excel (see Figure II.E.7) to examine the first 11 of these limiting probabilities for our example, in which $\mu = 3$ (since $\lambda = 24$ (per day) and $t = 1/8$ (of a day)). The remaining possibilities have low probabilities.

Figure II.E.7: [Calculating the probabilities for the Poisson distribution in Excel](#)

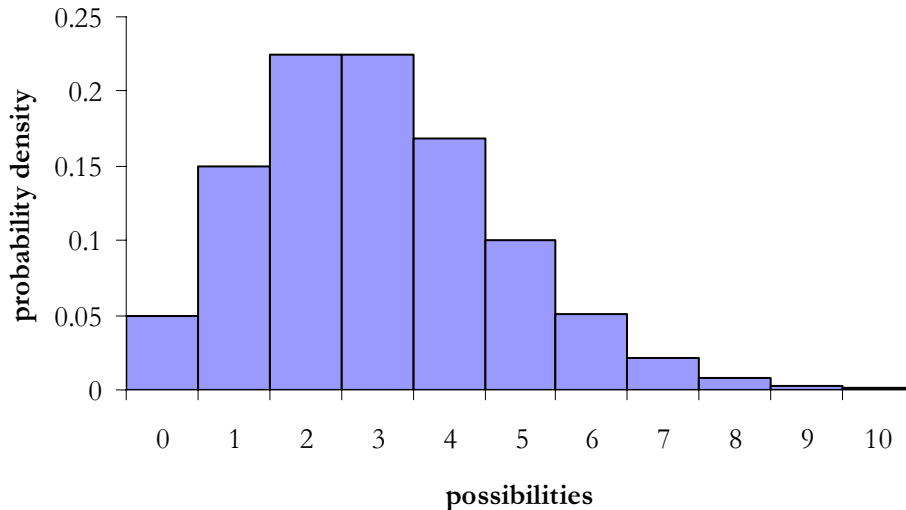


⁵ Siméon-Denis Poisson (1781–1840).

⁶ Recall that e raised to a power denotes the exponential function, also written ‘exp’ – see for instance Section II.C.1.4.10.

The histogram of these probabilities (Figure II.E.8) shows that the probability distribution is skewed to the right (i.e., that it is positively skewed).

Figure II.E.8: Histogram of probabilities for the Poisson distribution of Figure II.E.7



We will further illustrate the Poisson distribution by investigating the probability of there being more than three jumps in excess of 1% in the FTSE 100 index in the next six-month period.

Analysis of daily data of the FTSE 100 index from 3 January 1984 to 3 April 1992 shows that over that time the average number of daily changes in excess of 1% occurring in a six-month period was about 5. The number of such jumps may therefore be modelled as Poisson(5).

If $X \sim \text{Poisson}(5)$ then

$$P(X = 0) = \frac{5^0 e^{-5}}{0!} = e^{-5} \approx 0.0067$$

$$P(X = 1) = \frac{5^1 e^{-5}}{1!} = 5e^{-5} \approx 0.0337$$

$$P(X = 2) = \frac{5^2 e^{-5}}{2!} = 12.5e^{-5} \approx 0.0842$$

$$P(X = 3) = \frac{5^3 e^{-5}}{3!} = \frac{125}{6} e^{-5} \approx 0.1404$$

So the probability of there being more than three such jumps is $1 - (0.0067 + 0.0337 + 0.0842 + 0.1404) = 0.7350$.

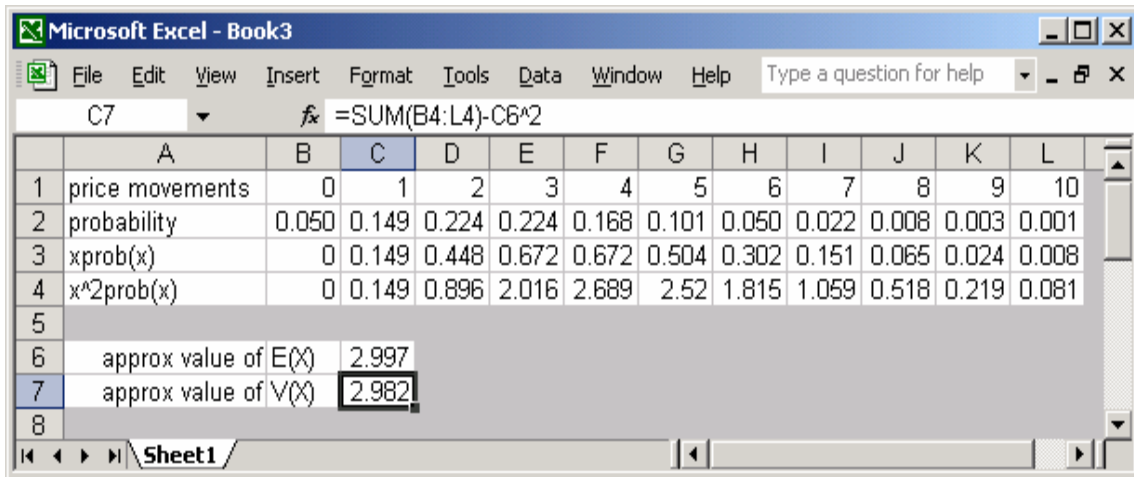
II.E.4.2.2 Expectation and Variance

If $X \sim \text{Poisson}(\mu)$ then:

$$E(X) = \mu; \quad \text{Var}(X) = \mu; \quad \sigma_X = \sqrt{\mu}. \quad (\text{II.E.26})$$

Since there are an infinite number of possibilities for X , we cannot demonstrate these results in the same way that we did with the binomial distribution, but we can do so approximately. The calculations are shown in Figure II.E.9.

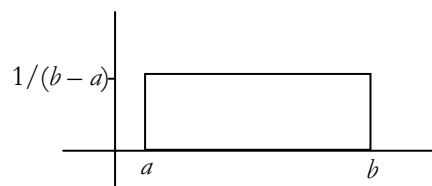
Figure II.E.9: Calculating the expectation and variance of a Poisson distribution in Excel



II.E.4.3 The Uniform Continuous Distribution

The pdf for the uniform continuous distribution is shown in Figure II.E.10.

Figure II.E.10: Probability density function of the Uniform(a, b) distribution



If $U \sim \text{Uniform}(a, b)$, then

$$E(U) = \frac{a+b}{2} \quad \text{and} \quad \text{Var}(U) = \frac{(b-a)^2}{12}.$$

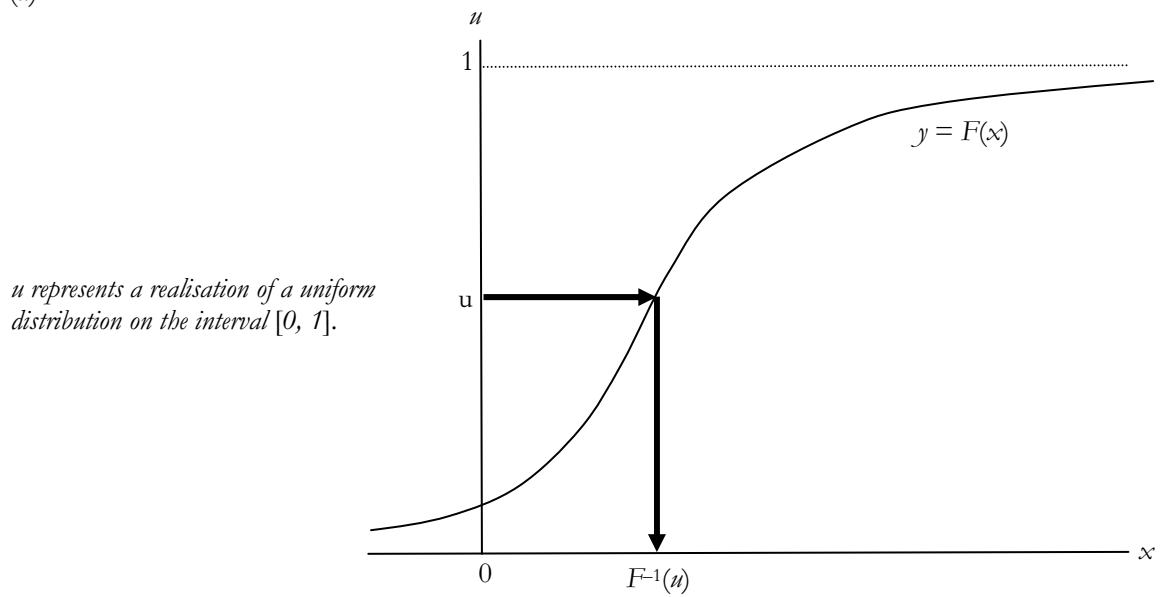
The uniform continuous distribution is used extensively within simulation applications (see Section II.D.4.2). This is because the method for choosing a random sample from a continuous

distribution is to apply the inverse of the cdf (Section II.E.2.3) to a random sample from a uniform distribution on the interval $[0, 1]$, denoted by u .

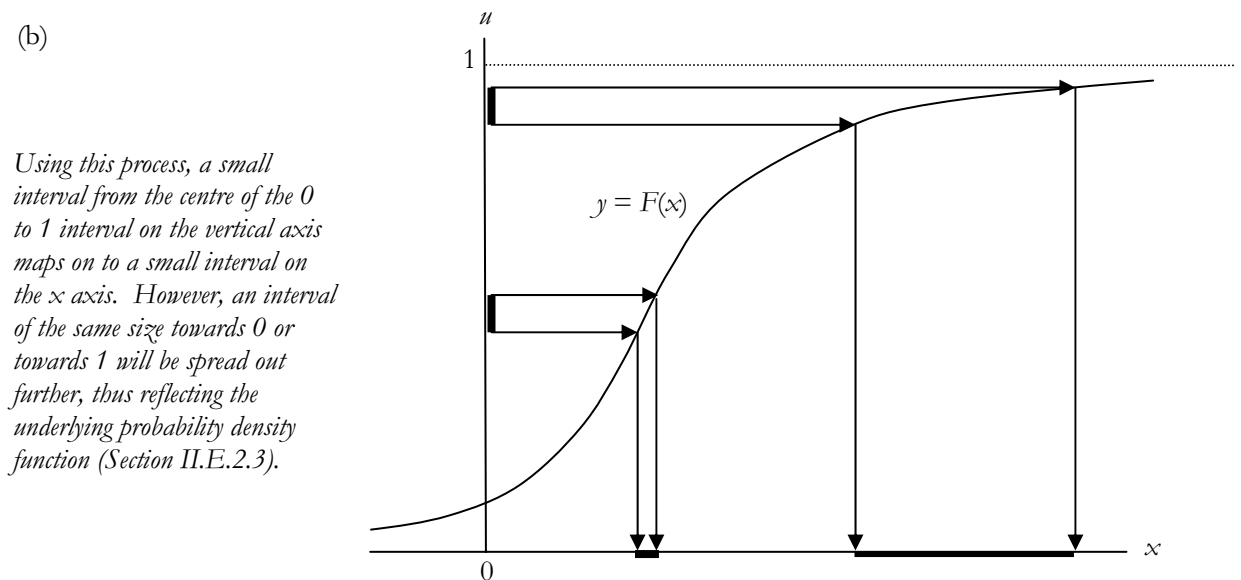
The process is illustrated in Figures II.E.11. Part (a) shows how to transform u into a realisation of the random variable X with cdf $F(x)$. Part (b) shows how the process captures and mirrors the probability density $f(x)$.

Figure II.E.11: Using the uniform distribution in simulations

(a)



(b)



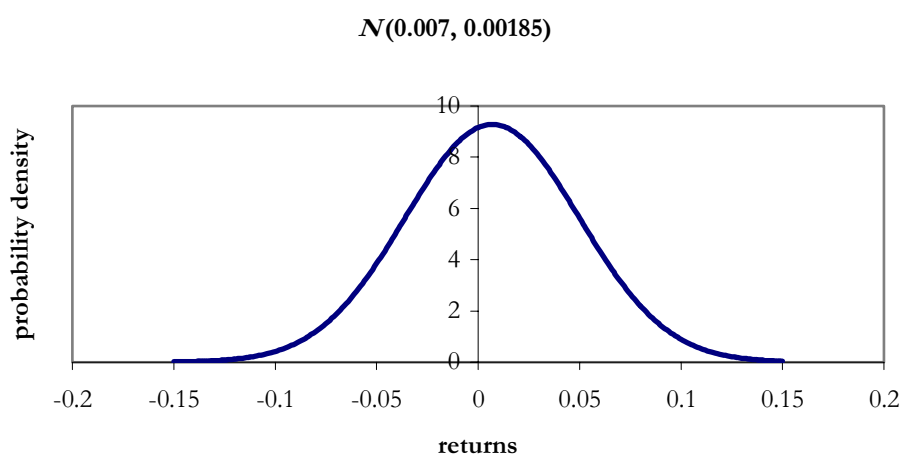
II.E.4.4 The Normal Distribution

The *normal* or *Gaussian distribution* is the most widely used probability distribution in finance. This is because of a result known as the *Central Limit Theorem*, which says that, for any distribution, the mean of a large number of independent realisations of that distribution is approximately normally distributed.

II.E.4.4.1 Normal Curves

The normal distribution is a continuous distribution, but is widely used as an approximate model for many discrete random variables. The graph of its pdf is a symmetrical bell shape, as shown in Figure II.E.12.

Figure II.E.12: [The normal pdf](#)



The distribution is completely defined by its mean and standard deviation. The mean indicates the position of the centre of the bell, and the standard deviation indicates how spread out the bell is. The particular normal distribution shown in Figure II.E.12 represents the continuously compounded monthly returns to the MSCI World Equity Index between 1980 and 2003.⁷ These were approximately normally distributed with a mean of 0.007 (0.7%) and a standard deviation of 0.043 (4.3%), and thus a variance of $0.043^2 = 0.00185$. This distribution is referred to as $N(0.007, 0.00185)$.

Note the symmetry of the curve and also how the tails gradually approach (although they never actually touch) the horizontal axis. They are said to asymptotically approach zero. As a consequence, the tails of the distribution are very small, and therefore the probability of extreme events is very small under a normal distribution.

⁷ These are computed as $\ln\left(\frac{\text{price at end of month } n+1}{\text{price at end of month } n}\right)$. (See section II.E.4.5 below.)

If a variable is normally distributed, 68.27% of the observations will fall within plus or minus one standard deviation from the mean. Moreover, 95.45% of the observations will fall within plus or minus two standard deviations, and 99.73% of the observations will fall within plus or minus three standard deviations from the mean. Therefore we can say in this example, with 95.45% confidence, that the returns for any given month are between $0.007 - 2 \times 0.043$ and $0.007 + 2 \times 0.043$, that is, between -0.079 and 0.093 .

The equation for the normal probability density function is

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (\text{II.E.27})$$

where μ is the mean of the distribution, σ is the standard deviation, $\pi \approx 3.14159$ and $e \approx 2.71828$.

II.E.4.4.2 The Standard Normal Probability Density Function

A *standard normal* variable, usually denoted by z , is one that has a mean of zero and a standard deviation of one. A realisation x of a normally distributed random variable X may be standardised by subtracting the mean and dividing by the standard deviation:

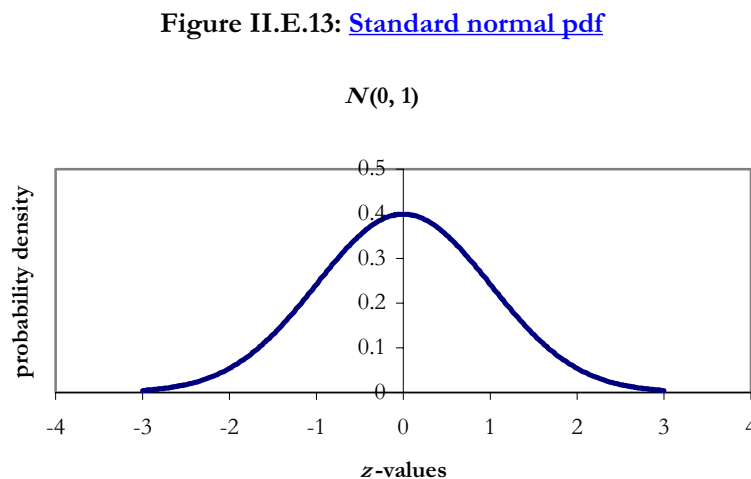
$$z = \frac{x - \mu}{\sigma}. \quad (\text{II.E.28})$$

The standardized variable, z , in (II.E.28) has the standard normal distribution, $N(0, 1)$, that is, it has a mean of zero and a variance (and thus a standard deviation) of one.

The equation for the standard normal probability density function is

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}. \quad (\text{II.E.29})$$

The graph of (II.E.29) is shown in Figure II.E.13.



II.E.4.4.3 Finding Areas under a Normal Curve Using Excel

Before packages such as Excel were available, normal probabilities were calculated using tables of the standard normal distribution. The upper and/or lower probabilities were standardised, and the probability could then be looked up in tables, since $P(x_1 < X < x_2) = P(z_1 < Z < z_2)$, where z_1 is the standardised value of x_1 and z_2 is the standardised value of x_2 .

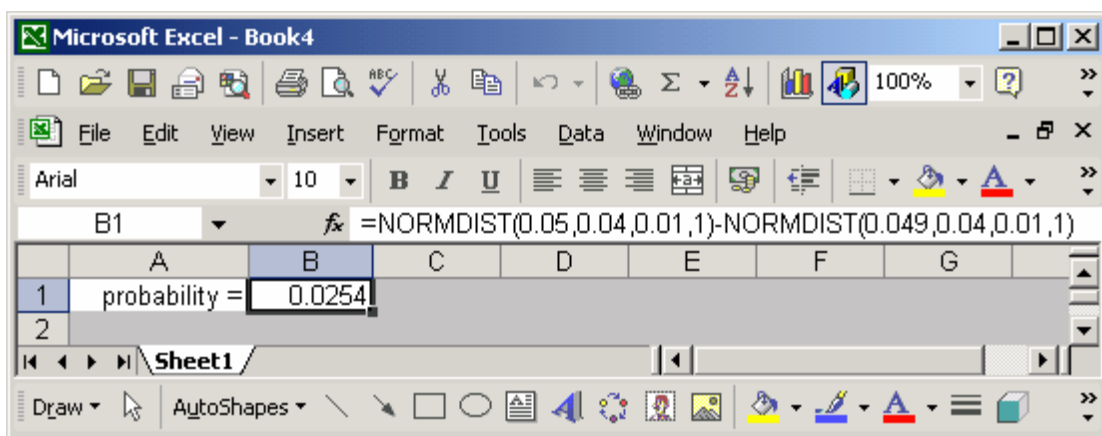
Excel returns the required probabilities through the functions NORMSDIST (for the standard normal distribution) and NORMDIST (for the non-standard normal). NORMDIST can also be used to give the height of the pdf. This was used to produce figures II.E.12 and II.E.13. That is, NORMDIST(0, 0.007, 0.00185, 0) gives the value (X) which is the y -intercept in Figure II.E.12, or the value assuming that the return is equal to zero. Note that the final zero within the parentheses indicates that the output will be the height of the pdf itself rather than the cdf. For calculating the cdf (i.e., the area under the curve to the left of the designated value) this indicator is set to one.

Suppose, for instance, that we wish to know the probability of a given asset, which is assumed to have normally distributed returns, providing a return of between 4.9% and 5%. The mean of the returns on the asset to date is 4%, and the standard deviation is 1%. NORMDIST(0.049, 0.04, 0.001, true) gives the probability of the returns being less than 4.9%, so the required probability is given by:

$$\text{NORMDIST}(0.05, 0.04, 0.001, 1) - \text{NORMDIST}(0.049, 0.04, 0.001, 1).$$

Here ‘1’ stands for ‘true’; it indicates that the cdf is required. The result is shown in Figure II.E.14.

Figure II.E.14: $P(0.049 < X < 0.05)$ for $X \sim N(0.04, 0.001)$



Note: When modelling a discrete distribution with a normal approximation, the process of counting probabilities from halfway between possibilities is known as ‘applying a continuity correction’. For the most part this is not an issue in mathematical finance.

II.E.4.5 The Lognormal Probability Distribution

A variable is said to be lognormally distributed if the natural logarithm of the variable is normally distributed.

Consider the continuously compounded returns to an asset over a period from time $t = 0$ to time $t = \tau$. The continuously compounded returns are calculated as $\ln(S(\tau)/S(0))$, that is, the log of the price relative. They are sometimes referred to as log returns. Under the standard model for asset prices these are taken to be normally distributed, so that the price relative itself, $S(\tau)/S(0)$ is lognormally distributed.

There are strong mathematical arguments which lead to the assumption of normality for continuously compounded returns. For a less formal but appealing argument, consider splitting the time interval up into small sub-intervals. Then

$$\frac{S(\tau)}{S(0)} = \frac{S(\delta t)}{S(0)} \times \frac{S(2\delta t)}{S(\delta t)} \times \frac{S(3\delta t)}{S(2\delta t)} \times \dots \times \frac{S(\tau)}{S(\tau - \delta t)}.$$

Taking logs gives

$$\ln\left(\frac{S(\tau)}{S(0)}\right) = \ln\left(\frac{S(\delta t)}{S(0)}\right) + \ln\left(\frac{S(2\delta t)}{S(\delta t)}\right) + \ln\left(\frac{S(3\delta t)}{S(2\delta t)}\right) + \dots + \ln\left(\frac{S(\tau)}{S(\tau - \delta t)}\right).$$

This is now an additive process, and providing that we assume independence across small time intervals, the Central Limit Theorem (see Section II.E.4.4) implies that $\ln(S(\tau)/S(0))$ should be normal.

This is an intuitively appealing model of the distribution of asset prices relatives because, if the price rises, the price relative will be greater than one, whereas if the price falls, the price relative will be less than one, but it cannot be negative. This also applies to asset prices, because the asset cannot have a negative value, and to equity prices, since liability is limited. We will return to this point later.

The lognormal probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2} \quad (\text{II.E.30})$$

where μ is the mean of the underlying normal distribution and σ is the standard deviation.

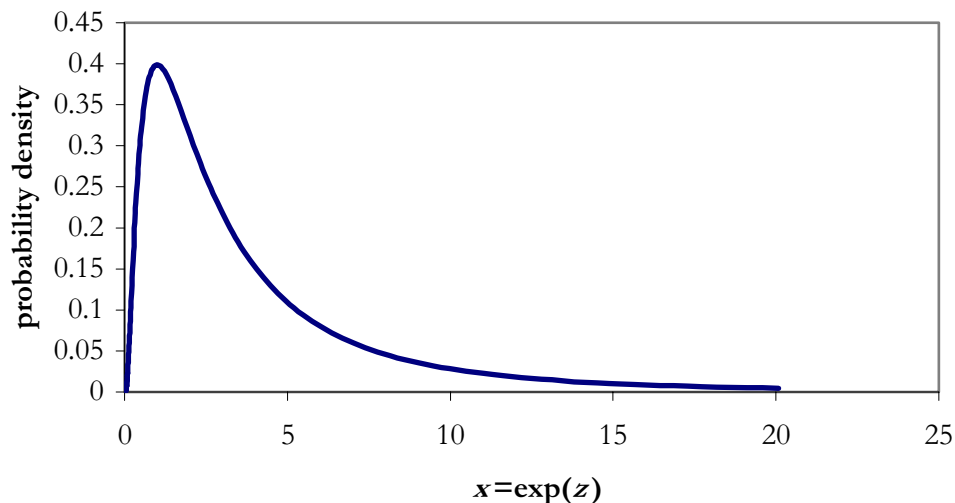
II.E.4.5.1 Lognormal Curves

The curve for the lognormal pdf is right-skewed, and bounded to the left – only positive values are possible. To see this consider Figure II.E.15, which shows the distribution of X when $\ln(X)$ is standard normal.

The lognormal distribution is skewed to the right and has no negative numbers. This shows its usefulness as a model of security price relatives, since price relatives cannot fall below zero, but a few observed values could be very high.

Figure II.E.15: [Lognormal distribution](#)

$$\ln(X) \sim N(0, 1)$$



II.E.4.5.2 The Lognormal Distribution Applied to Asset Prices

Consider again the continuously compounded returns to an asset over a period from time $t = 0$ to time $t = \tau$. Assume $\ln(S(\tau)/S(0)) \sim N(\mu, \sigma^2)$. (The values of μ and σ will depend upon the asset in question and the time period under consideration. Deriving expressions for them is a fundamental prerequisite to valuing derivative securities). We can re-express this as

$$\ln(S(\tau)) - \ln(S(0)) \sim N(\mu, \sigma^2),$$

or as

$$\ln(S(\tau)) \sim N(\ln(S(0)) + \mu, \sigma^2).$$

Note the fundamentally different nature of $S(\tau)$ and $S(0)$ in this analysis. $S(0)$ is the price now, and is a fixed, known number. $S(\tau)$ is the price in the future, and is a random variable. The above shows that the future price of the asset is also lognormally distributed. Furthermore, if

$$X \sim N(\ln(S(0)) + \mu, \sigma^2),$$

then

$$S(\tau) = S(0)\exp(x).$$

II.E.4.5.3 The Mean and Variance of the Lognormal Distribution

If $\ln(X) \sim N(\mu, \sigma^2)$, then

$$E(X) = \exp\left(\mu + \frac{\sigma^2}{2}\right), \quad (\text{II.E.31})$$

$$\text{Var}(X) = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1) \quad (\text{II.E.32})$$

These formulae are very useful indeed for many risk management applications. For example, in operational risk assessment the severity density is often assumed to be lognormal (see Chapter III.C.3). With data on the severity, X , one can use equations (II.E.31) and (II.E.32) to derive the parameters of the lognormal density function (II.E.30).

II.E.4.5.4 Application of the Lognormal Distribution to Future Asset Prices [not in PRM exam]

In the Black–Scholes theory of derivative pricing it is shown that

$$\ln\left(\frac{S(\tau)}{S(0)}\right) \sim N\left(\left(\alpha - \frac{\sigma^2}{2}\right)\tau, \sigma^2\tau\right),$$

where α is the *drift* rate and σ is the *volatility* of the underlying. The ‘drift’ is the growth rate in the expected asset price. The ‘volatility’ determines the random movements of the asset price around its expected value. To see that α is the drift, apply (II.E.31) to see that

$$E\left(\frac{S(\tau)}{S(0)}\right) = \exp\left(\left(\alpha - \frac{\sigma^2}{2}\right)\tau + \frac{\sigma^2\tau}{2}\right) = \exp(\alpha\tau).$$

Thus $E(S(\tau)) = S(0)e^{\alpha\tau}$, corresponding to our expectation of growth in the asset price with continuous rate α .

Suppose now that we need to know the probability of $S(\tau)$ being below some arbitrary number, K , in τ years’ time. This is the same as the $P(\ln(S(\tau)) < \ln(K))$. Since S is lognormally distributed, $\ln(S(\tau))$ will be normally distributed, so we can use a normal probability distribution to give us the probability we want.

Now we have that

$$\ln(S(\tau)) \sim N\left(\ln(S(0)) + \left(\alpha - \frac{\sigma^2}{2}\right)\tau, \sigma^2\tau\right).$$

So if Z is a standard normal variable then

$$P(\ln(S(\tau)) < \ln(K)) = P(Z < M),$$

where M is the standardised value of $\ln(K)$ given by

$$M = \frac{\ln(K) - \ln(S(0)) - \left(\alpha - \frac{\sigma^2}{2}\right)\tau}{\sigma\sqrt{\tau}} = \frac{-\ln\left(\frac{S(0)}{K}\right) - \left(\alpha - \frac{\sigma^2}{2}\right)\tau}{\sigma\sqrt{\tau}}.$$

Thus the required probability is given by

$$P(S(\tau) < K) = N\left(\frac{-\ln\left(\frac{S(0)}{K}\right) - \left(\alpha - \frac{\sigma^2}{2}\right)\tau}{\sigma\sqrt{\tau}}\right),$$

where N represents the standard normal cdf.

This will be familiar to you from Section I.A.8.7, because it is equivalent to $N(-d_2)$ in the Black–Scholes–Merton put option pricing model, except that the risk free return r of BSM is replaced by the expected return of the asset, α , using risk neutrality.

Example II.E.2:

Suppose that the underlying is an index with present price $S(0) = 4500$. Suppose that the strike of the option $K = 45,000$ and that the risk-free rate $r = \alpha = 3.5\%$. Suppose that τ (the maturity of the option) = 5 years and that the volatility of the index price is 17%. Then

$$P(S(5) < 4500) = N\left[\frac{-\ln\left(\frac{4500}{4500}\right) - 5(0.035 - 0.5 \times 0.17^2)}{0.17\sqrt{5}}\right] = 0.3935.$$

Thus the probability of the index having fallen in price in 5 years' time is 39.35%. Consequently there is a probability of $1 - 0.3935 = 0.6065$ that the index will be above 4500 at expiry.

II.E.4.6 Student's t Distribution⁸

Student's t distribution, also known simply as the t distribution, is a continuous distribution, originally developed to model the mean of small samples from a normal distribution when the variance is not known. It is most frequently used as a sampling distribution and for the creation

⁸ Developed by W. Gosset (1876–1937) who wrote under the pseudonym ‘A student of statistics’.

of confidence intervals for samples of less than 30. For large samples the Student t distribution converges to the normal distribution.

For small samples, the Student's t distribution has a higher level of kurtosis and has more weight in the tails. Consequently, it is sometimes used empirically as a model for the distribution of returns to financial assets. The shape of the t distribution is a function, amongst other things, of the number of degrees of freedom.

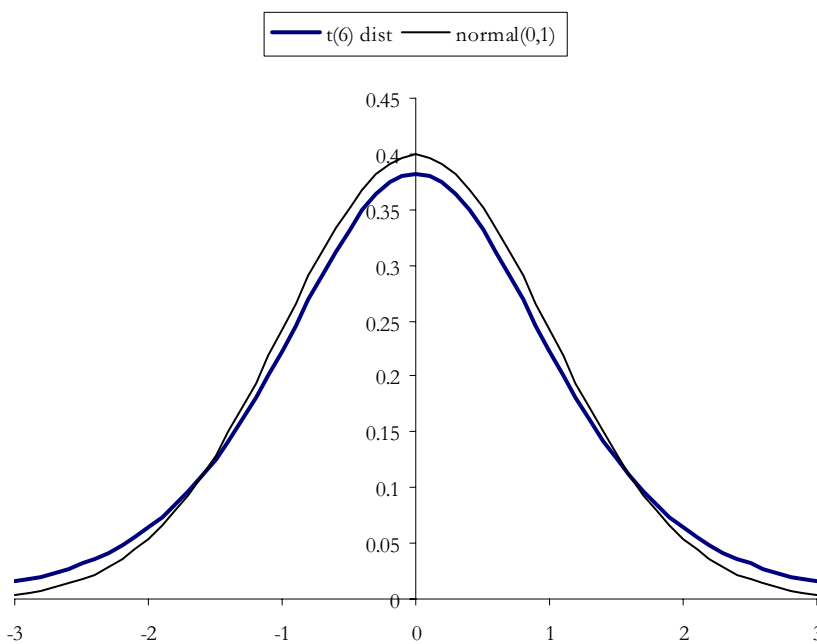
In ‘unconditional’ returns densities, the number of degrees of freedom is one less than the sample size. So with long runs of returns data, there will be many degrees of freedom, and the t distribution becomes indistinguishable from the normal distribution. Sometimes a small number of degrees of freedom has to be used, the number chosen being that which empirically gives the best fit to the data.

The density function is the formidable expression:

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi(\nu-2)\sigma^2}} \left(1 + \frac{(X-\mu)^2}{(\nu-2)\sigma^2}\right)^{-\frac{(\nu+1)}{2}} \quad (\text{II.E.33})$$

where Γ is the Gamma function and ν is the number of degrees of freedom. To evaluate Γ requires numerical integration unless the argument happens to be a positive integer when $\Gamma(n) = n!$ (Note that $\Gamma(0) = 1$, further justifying the use of $0! = 1$ in equation (II.E.21).)

Figure II.E.16: Comparing the $t(6)$ and $N(0,1)$ distributions



In Figure II.E.16 the Student t distribution with 6 degrees of freedom is juxtaposed with the standard normal distribution. The heavier tails of the t distribution are clearly shown, as is the fact that it unfortunately lacks the oft-required higher peak. Thus alternative heavy-tailed distributions are sometimes employed, such as normal mixture distributions. For further information about financial applications of normal mixture distributions, see www.ismacentre.rdg.ac.uk/alexander.

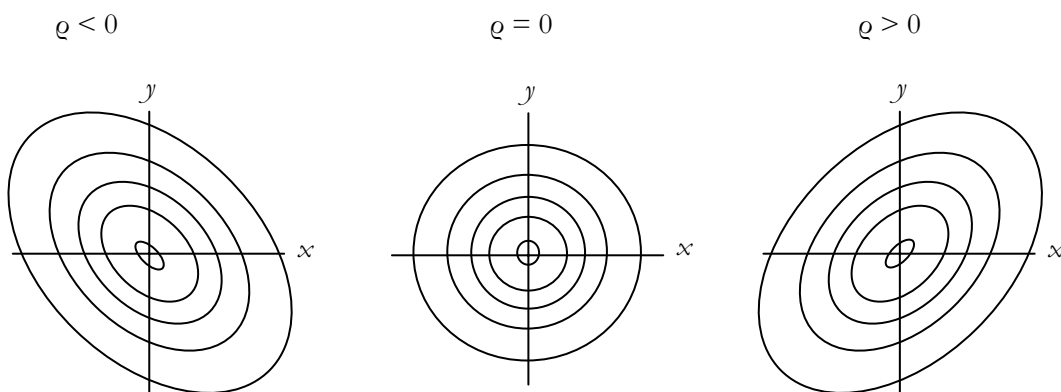
II.E.4.7 The Bivariate Normal Distribution

We finish by giving the probability density function of the *bivariate normal* distribution, plus a few words of commentary on it. This is an important example of a continuous bivariate distribution. Its risk management applications include the CreditMetrics methodology for estimating the credit loss distribution of a portfolio (see Chapter III.B.5).

You are already familiar with a bivariate distribution, albeit for *discrete* random variables (see Section III.E.3.1). A function of two variables requires three dimensions for its graph – x and y axes on the ‘floor’ and a vertical axis for values. Viewed in three dimensions, the joint density for discrete random variables is like a set of boxes. But two continuous random variables have a joint density that is like a *mountain*. Probabilities are represented by *volumes underneath the surface*. Two-dimensional projections can be drawn to illustrate this, but it is easier to draw contour diagrams, which we are all used to through reading maps.

The *bivariate normal density* has circular contours when the two variables have correlation $\rho = 0$, and elliptical contours otherwise, tending to straight lines for $\rho \rightarrow 1$ or $\rho \rightarrow -1$ (Figure III.E.17).

Figure III.E.17: Contours of three bivariate normal densities



Note that the ellipses are not necessarily inclined at 45 degrees. In fact, the major axis has gradient $\rho \frac{\sigma_Y}{\sigma_X}$ (see equation (II.E.34) below). Furthermore the centres are at the individual means, (μ_X, μ_Y) , not necessarily at $(0, 0)$ as illustrated

Formally, we say that X and Y have a bivariate normal distribution if their joint density function is:

$$f(x,y) = \frac{\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \quad \text{(II.E.34)}$$

where ρ is the correlation coefficient.

If X and Y have a bivariate normal distribution then the conditional density of Y given that X has value x is normal with mean

$$\mu_{Y|X=x} = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X) \quad \text{(II.E.35)}$$

and variance

$$\sigma_{Y|X=x}^2 = \sigma_Y^2(1 - \rho^2). \quad \text{(II.E.36)}$$

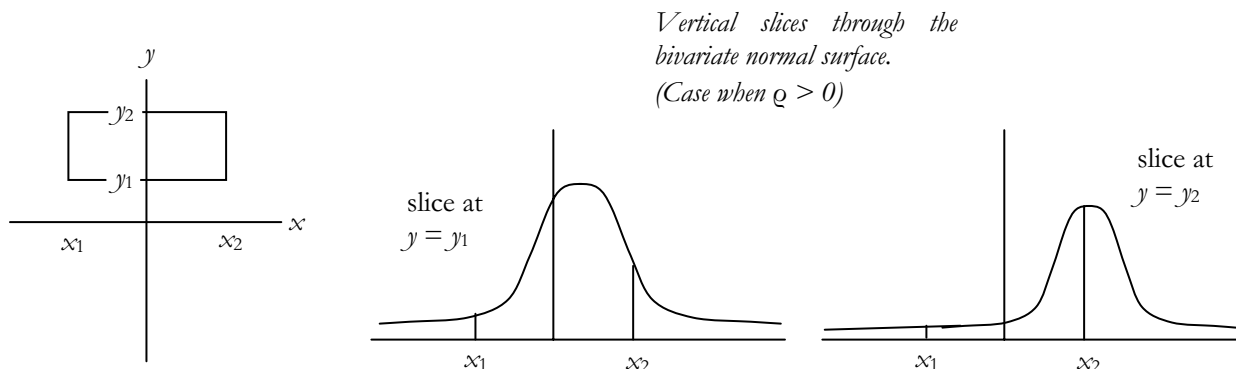
The symmetric result for ‘ X given that Y has value y ’ follows by swapping x s and y s. These are important results because they lie at the heart of regression analysis.

It follows from our remarks above that if X and Y are distributed bivariate normal, then

$$P((x_1 < X < x_2) \text{ and } (y_1 < Y < y_2))$$

is given by the volume above the rectangle and below the surface (see Figure II.E.18).

Figure II.E.18: The probability $P((x_1 < X < x_2) \text{ and } (y_1 < Y < y_2))$



II.F Regression Analysis in Finance

Keith Parramore and Terry Watsham¹

Regression analysis is widely used in finance to test mathematical models of the relationship between one variable (known as the dependent variable), and one or more independent variables. The dependent variable is usually referred to as the Y variable, and the independent variable(s), also referred to as *regressors* or *explanatory variables*, are referred to as the X variable(s).² Examples of the applications of regression in finance are:

- testing of the well-known capital asset pricing model developed by Sharpe (1964).
- the regression of an asset return on the return to an index representing the market, to estimate the sensitivity of the asset returns to variations in the market return. We will illustrate this process later in this chapter.
- the regression of portfolio returns on the returns of a futures or forward contract in order to determine the hedge ratio which minimises the variance of the returns on the hedged portfolio.
- the regression of the excess return of a portfolio (portfolio return minus risk-free rate of interest) on the excess return of a benchmark portfolio in order to assess the risk-adjusted performance.
- testing to determine whether asset prices follow a random walk. This usage is explained in more detail later in this chapter.

In this module we will explain two types of regression. They are *simple (univariate) regression* and *multiple regression*. Simple regression quantifies the dependence of the dependent variable on one explanatory variable, whereas multiple regression quantifies the dependence of the dependent variable on more than one explanatory variable. An example of simple regression would be where the returns of a single stock, say Hilton Hotels, are regressed on the returns of a broad index, say the FTSE 350 Leisure and Hotels index (henceforth the FTSE 350 L&H) to test the hypothesis that the returns to the Hilton shares are a function of the returns to the index. An example of multiple regression would be where the monthly returns to the Hilton shares are regressed upon the returns to the FTSE 350 L&H and the monthly change in air passenger traffic. This could test the hypothesis that the returns to the Hilton shares are influenced by *two factors*; in this example,

¹ Keith Parramore is Principal Lecturer in Mathematics at the University of Brighton and Terry Watsham is Principal Lecturer at the University of Brighton.

² The term 'regression' derives from the work of Francis Galton (1812–1911), who studied the data pairs (height of father, height of son). The line of best fit had a positive gradient, but a gradient which was less than 1, indicating a 'regression to the mean'.

positively influenced by movement in the index and positively influenced by increases in the volume of international travel as proxied by air passenger miles.

A distinction is made between *cross-section regression* and *time-series regression*. Cross-section regression tests the relationship between variables at a particular point in time. As an example of cross-section regression we may wish to measure the relationship between company size and the returns to investing in shares in the company. To do this we could collect data for many companies on share returns for a single period, say one year, and data on company size at the beginning of the same period. The returns data would represent the dependent variable, while the size data would be the independent variable. Thus the regression analysis describes the relationship between the variables at a single point in time. With time-series regression the data for each variable are collected over successive time periods. The Hilton Hotels example given earlier is a time-series regression because the data (observations of the stock returns, index returns and changes in air traffic) are collected over time. The regression analysis describes the relationship over the time period covered by the data. However, irrespective of whether the analysis is cross-sectional or time-series, the basic principles of regression analysis are the same.

II.F.1 Simple Linear Regression

II.F.1.1 The Model

In this section we apply regression analysis to the case of a simple linear³ relationship between the dependent (Y) variable and one independent (X) variable. By this we mean that the expected⁴ value of Y is given by $E(Y) = \alpha + \beta E(X)$ for some constant α (the intercept) and some constant β (the slope). The observed value of Y , denoted y , is then related to the observed value of X , denoted x , by

$$y = \alpha + \beta x + e. \quad (\text{II.F.1})$$

In this expression e is a realisation (i.e. a particular value) of a random variable ϵ with zero mean. This ϵ is referred to as the *error* variable, and a particular value e of the error is the *residual*. It reflects the fact that values of Y will be imperfectly described by values of X alone. There will be other factors not captured by the $\alpha + \beta x$ part of the model. However, in many cases these other factors are relatively unimportant. Whilst ϵ is often referred to as the error, this is not meant in the sense of a mistake, but rather in the sense that it represents the deviation of the actual value of Y from the expected value of Y .

In the simple regression models that are discussed in this chapter, the independent variable, X , is normally assumed to be a ‘deterministic’ variable, rather than a random or ‘stochastic’ variable.⁵

³ See the introduction to Chapter II.D for a full description of linearity.

⁴ See Sections II.E.2.5 and II.E.2.7.

⁵ In more advanced regression analysis the independent variables are also random variables.

This may not be very realistic, but the statistical theory is much simplified by this assumption. Another common assumption is that ϵ is normally distributed,⁶ i.e. $\epsilon \sim N(0, \sigma^2)$.

Now consider the two parameters in (II.F.1):

- The parameter α is the constant term that corresponds to the expected value of Y when X is zero. We also call an estimate of α the *intercept* because if the data on X and Y are plotted on a *scatter plot* as below, α corresponds to the point where the line along which the scatter of data observations lies crosses the Y axis.
- The parameter β is known as the regression coefficient. Again, if the data are plotted on a scatter plot then β is the slope of the line along which the scatter of data observations lies. It can be interpreted as indicating the expected change in the variable Y that is caused by a one-unit change in the value of X . The sign of β is positive if the dependent and the independent variables are positively correlated. The sign is negative if they are negatively correlated.

Below we will estimate and test the relationship between the returns on Hilton shares and the returns on the FTSE 350 L&H index. The data can be found in the Excel workbook II.F.xls. The returns on the Hilton shares constitute the dependent variable, Y , because we have hypothesised that its movements are influenced by (i.e. are dependent on) the returns of the FTSE 350 L&H index (represented by X).⁷ Thus we assume that the many other, random, influences are represented by the random variable ϵ .

Our model is:

$$R_{\text{Hilton}} = \alpha + \beta R_{\text{FTSE}} + \epsilon. \quad (\text{II.F.2})$$

It should be noted that the data tell us nothing about causality, nor does our model make any assumptions regarding causality. Our understanding of any causal link can only come from an *a priori* hypothesis. As we noted in the previous paragraph, the direction of any cause and effect (i.e. which is the dependent variable and which is the independent variable) is determined by our hypothesis. The hypothesis has to be developed independently of the regression model so that the regression analysis can validly support or not support the hypothesis. Regression analysis cannot ‘prove’ a hypothesis, it can only support it statistically or reject it.

⁶ See Section II.E.4.4.

⁷ It is important to note that the regression here relates to percentage returns, not prices. The reason is that for ordinary least-squares regression to be applied to time-series data we require stationarity. Briefly that means that the data exhibit a constant mean, a constant variance and that the covariance between observations is a function only of the time period between observations. For a brief discussion of this issue refer to Section II.F.8. For a fuller discussion refer to Chapter 7 of Watsham and Parramore (1997).

II.F.1.2 The Scatter Plot

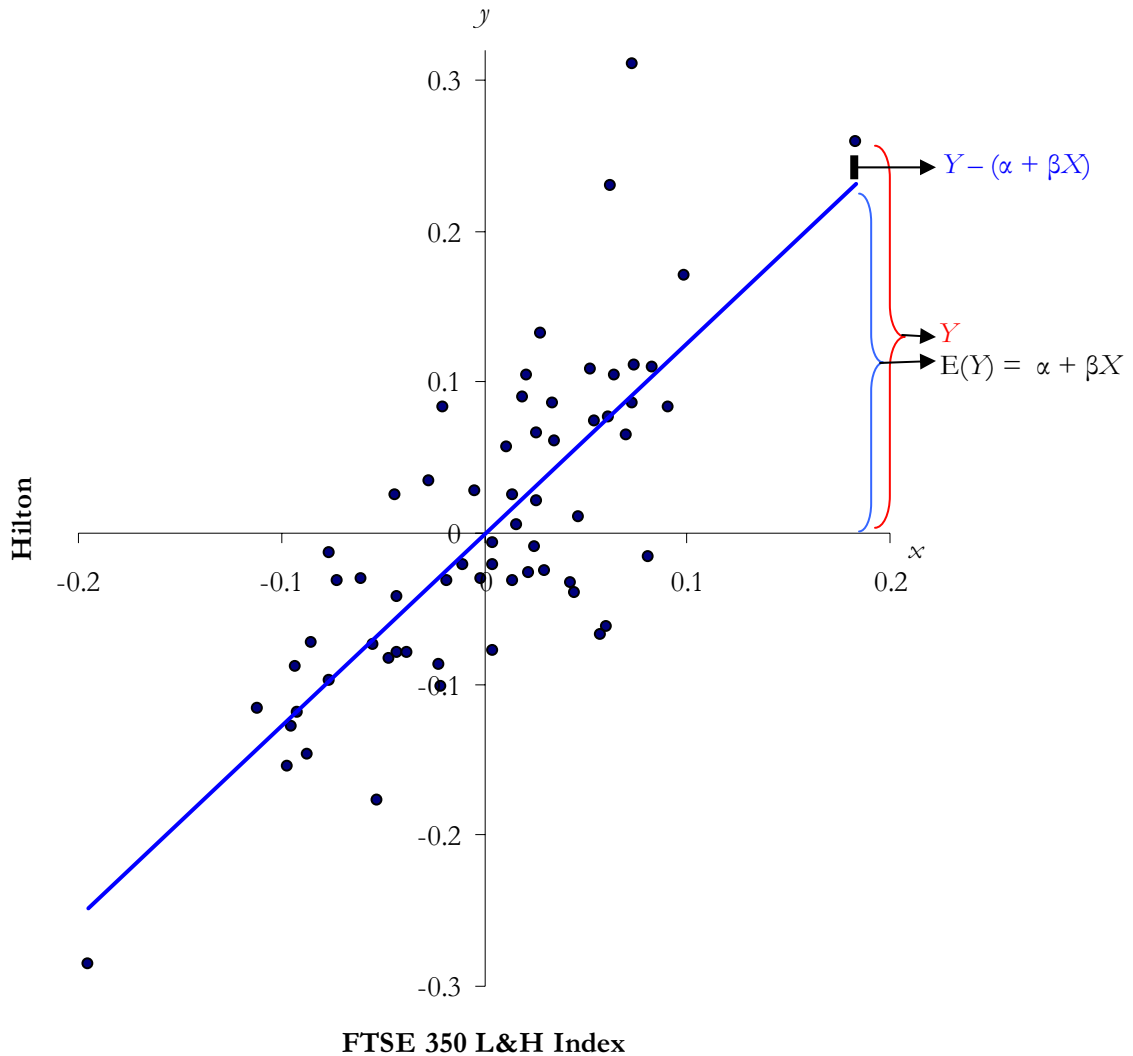
The first stage of our analysis is to plot a scatter plot of the returns data of the two variables. This consists of all points $(R_{\text{Hilton}}, R_{\text{FTSE}})$ in the plane. The time period is between the end of January 1999 and the end of April 2004 and the data are monthly returns. Looking at Figure II.F.1, we see a positive relationship between the Hilton returns and the FTSE 350 L&H index returns. But we need to measure what that relationship is, on average, over the time period of our data.

In *ordinary least-squares* (OLS)⁸ *regression* we calculate the linear relationship which minimises the sum of the squares of the residuals, that is, the sum of the squared differences between the values of Y predicted by our straight line and the actual observations on Y . This is shown in Figure II.F.1.

⁸ *Ordinary* least squares contrasts with further developments such as *weighted* least squares. We will only be considering ordinary least squares.

Figure II.F.1: [Scatter plot and regression line](#)

Returns to Hilton Group PLC v Returns to FTSE 350 Leisure & Hotels Index



II.F.1.3 Estimating the Parameters

The general form of the relationship between the dependent variable Y , and the independent variable X , in the simple linear regression model is

$$Y = \alpha + \beta X + \epsilon. \tag{II.F.3}$$

From the known value of X we can compute an expected value for Y , derived from estimates of α and β . Our estimates of the true values of α and β are computed from the data, and are labelled $\hat{\alpha}$ and $\hat{\beta}$, respectively. We can then compute the expected value of Y from

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X . \quad (\text{II.F.4})$$

Thus for each value x_i there is an observed value y_i and an estimated expected value, \hat{Y}_i , given by applying equation (II.F.4). The difference between y_i and \hat{Y}_i gives a value for the residual, e_i . OLS regression produces a straight line through the data that minimizes the sum of the squares of the e_i s, that is, minimizes

$$\sum_i (y_i - \hat{Y}_i)^2 .$$

For obvious reasons, this is known as the *residual sum of squares* (RSS). From equation (II.F.4) we can see that the residual sum of squares is also given by

$$\text{RSS} = \sum_i (y_i - (\alpha + \beta x_i))^2 . \quad (\text{II.F.5})$$

It is this that we wish to minimise. For those not used to ‘sigma’ notation it might help to write the expression out, thus:

$$\text{RSS} = (y_1 - (\alpha + \beta x_1))^2 + (y_2 - (\alpha + \beta x_2))^2 + (y_3 - (\alpha + \beta x_3))^2 + \dots + (y_n - (\alpha + \beta x_n))^2 . \quad (\text{II.F.6})$$

This should make it clear that, contrary to intuition, it is the α and β which are acting as variables in the expression. The x_i and y_i are fixed data observations. The α and β are for us to choose. The techniques required to enable us to make the best choice are covered in Chapter II.C, specifically in Sections II.C.5 and section II.C.7. They give:

$$\frac{\partial \text{RSS}}{\partial \alpha} = -2(y_1 - (\alpha + \beta x_1)) - 2(y_2 - (\alpha + \beta x_2)) - 2(y_3 - (\alpha + \beta x_3)) - \dots - 2(y_n - (\alpha + \beta x_n)),$$

$$\frac{\partial \text{RSS}}{\partial \beta} = -2x_1(y_1 - (\alpha + \beta x_1)) - 2x_2(y_2 - (\alpha + \beta x_2)) - 2x_3(y_3 - (\alpha + \beta x_3)) - \dots - 2x_n(y_n - (\alpha + \beta x_n))$$

Using sigma notation to simplify these expressions, we get:

$$\frac{\partial \text{RSS}}{\partial \alpha} = -2 \left(\sum_i y_i - n\alpha - \beta \sum_i x_i \right) \quad \text{and} \quad \frac{\partial \text{RSS}}{\partial \beta} = -2 \left(\sum_i (x_i y_i) - \alpha \sum_i x_i - \beta \sum_i x_i^2 \right) . \quad (\text{II.F.7})$$

Putting $\frac{\partial \text{RSS}}{\partial \alpha} = 0$ and $\frac{\partial \text{RSS}}{\partial \beta} = 0$ gives what are known as the *normal equations*. Solving the normal

equations gives our OLS estimates $\hat{\alpha}$ and $\hat{\beta}$:

$$\hat{\beta} = \frac{\sum_i (x_i y_i) - \sum_i x_i \sum_i y_i / n}{\sum_i x_i^2 - \left(\sum_i x_i\right)^2 / n} = \frac{\text{cov}(X, Y)}{\text{var}(X)}, \quad (\text{II.F.8})$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}. \quad (\text{II.F.9})$$

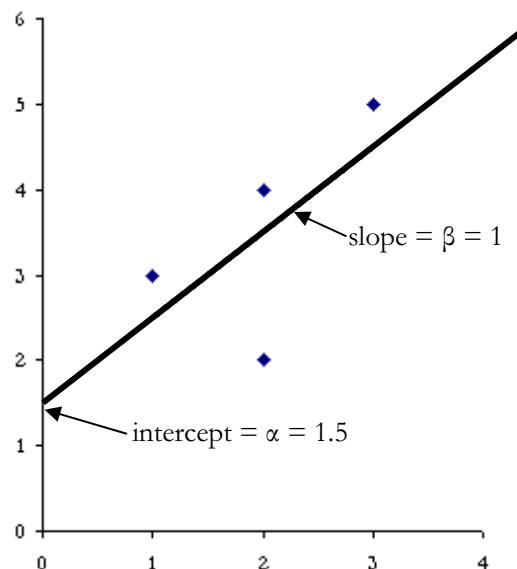
Example II.F.1

In this simple example the parameters are calculated from equations (II.F.8) and (II.F.9).

	<i>x</i>	<i>y</i>	<i>x</i> ²	<i>xy</i>
	1	3	1	3
	2	2	4	4
	2	4	4	8
	3	5	9	15
Sums	8	14	18	30

$$\text{cov}(X, Y) = \frac{(30 - 4 \times 2 \times 3.5)}{3} = \frac{2}{3}$$

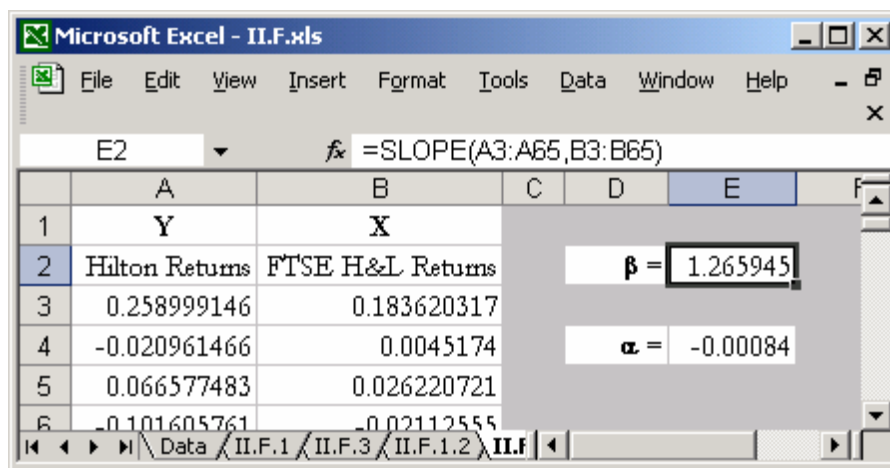
$$\text{var}(X) = \frac{(18 - 4 \times 2^2)}{3} = \frac{2}{3}$$



Example II.F.2: Hilton returns v. FTSE H&L returns

In this example the parameters are estimated using the Excel spreadsheet functions SLOPE and INTERCEPT (see Figure II.F.2).

Figure II.F.2: Calculation of α and β in Excel



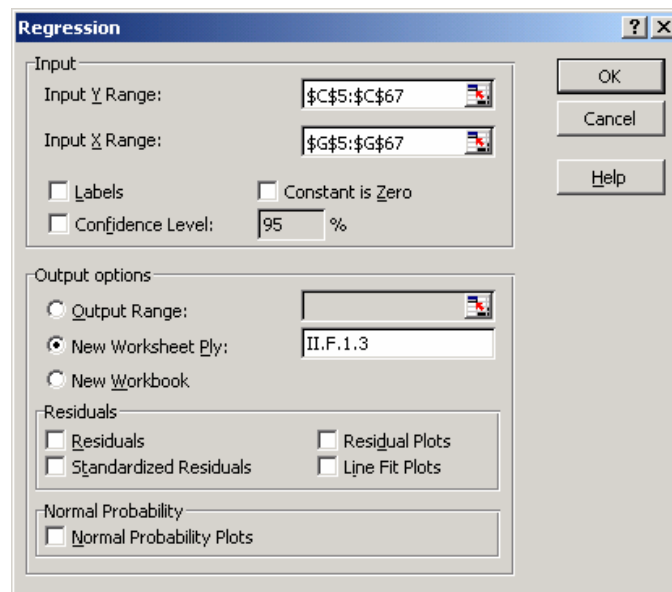
Note that the intercept here is virtually zero, as can also be seen in Figure II.F.1 where the regression line seems to cut through the origin. The estimated version of the model (II.F.2) is therefore:

$$E(R_{\text{Hilton}}) = 1.266E(R_{\text{FTSE}}).$$

Having computed our regression line, the *line of best fit*, we next have to consider how good the fit is. This is essentially a question concerning the nature of the error random variable, ϵ . However, there are associated questions such as those concerning the accuracy of our parameter estimates and the accuracy of any predictions which we might make using the model. To deal with some of these questions we need to develop confidence intervals and tests of hypotheses. We do that later in the chapter, but at this point it is appropriate that we show how to obtain a fuller analysis from Excel.

There are many statistical packages available that will carry out the full calculations. We will illustrate the application of the regression facility in Excel. The OLS regression tool in Excel is found in the Tools/Data Analysis/Regression (see Figure II.F.3).⁹ Here you are asked to define the Y values (in our case, the returns to the Hilton stock), as well as the X values (the returns to the FTSE 350 L&H Index). There are a number of other options available (such as showing residuals, saving the information on another worksheet, etc.). Note that in this case, we have refrained from defining an output range for the data. As a result, the regression results will be shown on a new worksheet.

Figure II.F.3: Regression dialog box in Excel



When you have finished, press OK. The results screen shown in Figure II.F.4 will appear.

Figure II.F.4: [Results screen for Hilton returns regression](#)

SUMMARY OUTPUT						
Regression Statistics						
Multiple R		0.785199504				
R Square		0.616538261				
Adjusted R Square		0.610252003				
Standard Error		0.065628685				
Observations		63				
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	0.422430467	0.422430467	98.07714868	2.58135E-14	
Residual	61	0.26273458	0.004307124			
Total	62	0.685165048				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-0.00084233	0.008272424	-0.101823877	0.919230447	-0.017384069	0.015699408
X Variable 1	1.26594499	0.12782945	9.903390767	2.58135E-14	1.010334136	1.521555844

For this analysis, the following three fields are important (all marked with a green box):

R^2 is equal to 0.617 (see below);

α , the intercept, is estimated as -0.001 ;

and β , the slope, is estimated as 1.266 (denoted by 'X Variable 1').

The intercept and slope correspond to our earlier computations. The R^2 is related to the ANOVA table below the regression statistics. ANOVA stands for 'analysis of variance'. The total variance in the dependent variable is given by the variation in Y about its mean. We call this the *total sum of squares* (TSS), defined as $\sum_i (y_i - \bar{y})^2$. Ideally, the regression model will *explain* much of the

variation in Y . The amount of variation explained by the model is $\sum_i (y_i - \bar{y})^2 - \sum_i (y_i - \hat{Y}_i)^2$,

and this is known as the *explained sum of squares* (ESS) or the *regression sum of squares* in some statistical packages, including Excel. Thus, in this example TSS = 0.685165 and ESS = 0.42243. The explanatory variable (returns to FTSE 350 L&H) explains a proportion $0.42243/0.68517 = 0.61653$, that is, about 62% of the variability of the data.

⁹ If you do not have the Data Analysis option under Tools, you need to add in this option. To do so, choose the option Tools/Add Ins, and then mark the two add-ins Analysis ToolPak and Analysis ToolPak – VBA. Click OK to complete this task.

The proportion of the variability in Y that is accounted for by the model is called R^2 :

$$R^2 = \frac{\text{the regression sum of squares}}{\text{the total sum of squares}} = \frac{\text{ESS}}{\text{TSS}}.$$

Its value gives an instant feel for how effective the model is. In this case the output shows immediately that R^2 is 0.61653, or you can calculate it using ESS and TSS as above.

Note that it is always the case that $\text{TSS} = \text{ESS} + \text{RSS}$ as in the ANOVA table above. This can be written as

$$R^2 = 1 - (\text{RSS}/\text{TSS}).$$

Hence R^2 has a value ranging from zero, where X has no influence upon Y (and $\text{RSS} = \text{TSS}$), to one, where all of the variation in Y is explained by the variation in X (and $\text{RSS} = 0$). It can be shown that, in the case of simple linear regression, the ratio ESS/TSS is equal to the square of the sample correlation coefficient (see Section II.B.6.3), hence R^2 is also known as the *coefficient of determination*. It gives a useful feel for the significance of the correlation coefficient – a correlation coefficient of 0.7 indicates that a linear model accounts for about 50% of the variability of the data, since $0.7^2 = 0.49$.

II.F.2 Multiple Linear Regression

II.F.2.1 The Model

Usually the behaviour of the dependent variable is best explained by more than one independent variable, used in combination. In these circumstances we apply *multiple regression*. Generalising the work on simple linear regression, the true relationship between the dependent random variable, Y , and the various deterministic independent variables, X_i ($i = 1, 2, 3, \dots, n$), is given by

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_n X_n + \epsilon. \quad (\text{II.F.10})$$

As in the case of simple linear regression, we do not know the true relationship and have to make estimates:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \dots + \hat{\beta}_n X_n \quad (\text{II.F.11})$$

The $\hat{\beta}$ s represent the partial derivatives of \hat{Y} with respect to the appropriate X s. For example,

$$\hat{\beta}_1 = \frac{\partial \hat{Y}}{\partial X_1}, \quad \hat{\beta}_2 = \frac{\partial \hat{Y}}{\partial X_2}, \quad \hat{\beta}_n = \frac{\partial \hat{Y}}{\partial X_n}. \quad (\text{II.F.12})$$

Multiple regression has been used to estimate multi-factor models of asset returns, including much testing of the arbitrage pricing theory developed by Ross (1976). It is also used in the development of many macro-economic models. Multiple regression is often used as an integral part of more sophisticated econometric techniques, particularly in time-series analysis, but these are beyond the scope of this module.

To illustrate multiple regression here, we will develop a ‘multi-factor’ model by adding another explanatory variable to the model of the returns to the Hilton shares. That variable is monthly data on airline passenger traffic. There are many computer packages that solve such a multiple regression problem; again we will use Excel.

II.F.2.2 Estimating the Parameters

As with simple regression we have two alternatives. We can use appropriate Excel functions such as SLOPE, INTERCEPT, FORECAST and STEYX. The advantage of this approach is that the outputs are *dynamic* – change the data and the outputs will change. Using the Analysis Toolpak is easier, but it is static – change the data and the analysis must be rerun.

In the next example we show how the returns of the Hilton shares can be estimated in dependence on the FTSE 350 L&H index and on the airline passenger traffic. In other words, a multiple regression is to be estimated, with the returns of the Hilton shares being the dependent variable and the returns on the FTSE 350 L&H index and the changes in airline passenger traffic being two independent variables. The estimation time period is again between the end of January 1999 and the end of April 2004.

Figure II.F.5: Regression dialog box for multiple regression

Regression

Input

Input Y Range:

Input X Range:

Labels Constant is Zero

Confidence Level: %

Output options

Output Range:

New Worksheet Ply:

New Workbook

Residuals

Residuals Residual Plots

Standardized Residuals Line Fit Plots

Normal Probability

Normal Probability Plots

OK
Cancel
Help

Figure II.F.6: Results screen for multivariate regression

	A	B	C	D	E	F	G	H	I	J	K
1	0.258999	0.18362	0.028681		SUMMARY OUTPUT						
2	-0.02096	0.004517	0.088251								
3	0.066577	0.026221	-0.01787		Regression Statistics						
4	-0.10161	-0.02113	-0.01256		Multiple R	0.79156677					
5	-0.08386	-0.04682	0.063924		R Square	0.62657794					
6	-0.02516	0.030265	0.020861		Adjusted R Square	0.61413054					
7	-0.03106	-0.07255	-0.02086		Standard Error	0.06530132					
8	-0.11936	-0.09235	-0.09531		Observations	63					
9	-0.1288	-0.09523	0.029981								
10	0.082734	0.091366	-0.00564		ANOVA						
11	-0.01626	0.080896	-0.06426			df	SS	MS	F	Significance F	
12	-0.1552	-0.09692	-0.03844		Regression	2	0.429309306	0.214665	50.33805	1.466E-13	
13	0.311332	0.073128	0.047446		Residual	60	0.255855741	0.004264			
14	0.230238	0.061933	0.110281		Total	62	0.685165048				
15	-0.07933	-0.04302	-0.00537								
16	-0.08718	-0.02174	0.005369			Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	-0.06265	0.060581	0.07229		Intercept	-0.00115571	0.008234857	-0.14034	0.888858	-0.0176279	0.0153165
18	-0.03064	-0.06038	-0.00625		X Variable 1	1.19781969	0.138039113	8.677393	3.46E-12	0.9217004	1.4739389
19	-0.03161	-0.01785	-0.02796		X Variable 2	0.16685759	0.131374276	1.270093	0.208955	-0.09593	0.4296452
20	-0.11664	-0.11187	-0.1045								

The analysis may require some data editing – Excel needs the independent variables to be in adjacent columns, so some copying, using Paste Special/by Value, may be needed. The Analysis Toolpak is used as shown in Figure II.F.5.

The green highlighted cells in Figure II.F.6 show that the intercept is -0.0012 , that β_1 is estimated as 1.1978 and that β_2 is estimated as 0.1669 . Therefore the estimated model is

$$\hat{Y} = -0.0012 + 1.1978X_1 + 0.1669X_2 .$$

This may be interpreted as follows: if X_2 is held constant, then the expectation is that a one-unit change in X_1 (FTSE 350 L&H) will cause a 1.198 change in \hat{Y} . Similarly, if X_1 is held constant, then the expectation is that a one-unit change in X_2 (airline passenger traffic) will cause \hat{Y} to change by 0.167 units. The value for R^2 is 0.6266, showing that the model explains about 63% of the variability in the Hilton equity returns – not, in this case, much of an improvement on the simple regression model.

II.F.3 Evaluating the Regression Model

The evaluation of the regression model is carried out in two phases. The first stage is the intuitive evaluation: does the model appear to support our hypothesis or not? Then the second stage rigorously analyses the validity of the results. For simplicity we will illustrate the interpretation with regard to the univariate regression example. However, the principles are equally applicable to multiple regression.

II.F.3.1 Intuitive Interpretation

The first stage of testing is to ensure that the intuitive interpretation of the results is compatible with the original hypothesis being tested. Particularly important here are:

1. the magnitude and sign of each regression parameter;
2. the magnitude of R^2 .

In our original hypothesis the returns to the Hilton shares are positively influenced by the returns to the FTSE 350 L&H index. So we would expect the sign of the FTSE 350 L&H index parameter to be positive, which it is. Also the R^2 of 0.61653 in Example II.F.2 indicates that about 62% of the variation in Y (the returns to the Hilton shares) is explained by the variation in X (the returns to the FTSE 350 L&H index).

II.F.3.2 Adjusted R^2

In multivariate regression, adding additional explanatory variables will cause the coefficient of determination to increase. Consequently, the coefficient of determination should be adjusted to take account of the number of independent variables. The adjusted R^2 , denoted by \bar{R}^2 , is calculated as

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n - 1}{n - 1 - k} \quad (\text{II.F.13})$$

where n is the number of observations and k is the number of independent regressors.

To illustrate this, recall the earlier R^2 of 0.62657 with two regressors (Section II.F.2.2). This gives:

$$\bar{R}^2 = 1 - \left[(1 - 0.62657794) \times \frac{62}{60} \right] = 0.61413054.$$

Using the adjusted R^2 helps to avoid falling into the trap of repeatedly introducing new variables in order to increase the fit. Introducing more variables may improve the fit, but may not improve the model. It is not worth gaining an insignificant increase in fit at a cost of extra complication. A good, parsimonious model is to be preferred to an overcomplicated model, possibly using inappropriate independent variables and giving only a slightly better fit. But in any case, the reader is cautioned against adding or deleting variables simply according to their influence upon R^2 or the adjusted R^2 . The rational basis for inclusion or deletion is the theory behind the model that is being tested.

II.F.3.3 Testing for Statistical Significance

The second stage is to test the statistical significance of the regression results. This is important because, as we noted earlier, probabilistic models only provide estimates of the regression coefficients. It is important therefore to test how representative these estimates are of the true coefficients. This is achieved by testing the statistical significance of the regression coefficients and the closeness of fit of the data to the estimated regression line. To carry out these tests we will use standard errors, t -statistics and p -values, all of which will be explained as we go along.

II.F.4 Confidence Intervals

The calculations we have covered so far give what are known as *point estimates* of the regression parameters. We know that these estimates are inaccurate to some degree, simply because they *are* estimates. We need to know the degree of confidence that we can place on these point estimates. This degree of confidence is articulated by the *confidence interval*, which is explained below.

II.F.4.1 Confidence Intervals for the Regression Parameters

In building the regression model we have made several assumptions about the error term(s). We have assumed that error terms are independent and identically distributed (i.i.d.). If we are prepared further to assume that the error terms are *normally* distributed, then we can begin to make inferences regarding the accuracy of our parameter estimates and of forecasts, etc.

Thus we shall assume the following:

1. The error random variable, ε , is normally distributed, with a mean of zero and a constant variance, σ^2 , that is, $\varepsilon \sim N(0, \sigma^2)$. Note that the assumption of constant variance is the *homoscedasticity* assumption.¹⁰
2. The successive error terms are independent of each other, which implies that the covariance between pairs of error terms is zero ($\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$). This is the assumption of no *autocorrelation*.

Under this assumption of normality it can be shown that our estimates $\hat{\alpha}$ and $\hat{\beta}$ are also normally distributed. Furthermore, the estimates are *unbiased*, that is, $E(\hat{\alpha}) = \alpha$ and $E(\hat{\beta}) = \beta$, and their standard deviations can be expressed in terms of the standard deviation of ε , the error random variable. These standard deviations, in the context of estimation, are known as *standard errors*.

We need to know how to use these statistics. We know, for instance, that $\hat{\beta}$ is normally distributed with mean β and variance $(\text{se}\hat{\beta})^2$, where $\text{se}\hat{\beta}$ is the standard error of $\hat{\beta}$, that is, $\hat{\beta} \sim N(\beta, (\text{se}\hat{\beta})^2)$. From the properties of the normal distribution this means that 95% of realisations of $\hat{\beta}$ will lie between $\beta - 1.96\text{se}\hat{\beta}$ and $\beta + 1.96\text{se}\hat{\beta}$, that is,

$$\text{Prob}(\beta - 1.96\text{se}\hat{\beta} < \hat{\beta} < \beta + 1.96\text{se}\hat{\beta}) = 0.95.$$

With a little algebra this translates into¹¹

$$\text{Prob}(\hat{\beta} - 1.96\text{se}\hat{\beta} < \beta < \hat{\beta} + 1.96\text{se}\hat{\beta}) = 0.95. \quad (\text{II.F.14})$$

Note that we do not know the standard error of ε , and therefore we do not know the standard error of $\hat{\beta}$, which depends on it. They can be estimated, but the details need not concern us, since Excel and other packages routinely compute and display estimates of the standard errors.

¹⁰ ‘Homoscedastic’ is Greek for constant (or same) scale.

¹¹ Note: This ‘little algebra’ produces a result which looks odd. It looks as if we are saying that there is a 95% chance that the parameter β lies between $\hat{\beta} - 1.96\text{se}\hat{\beta}$ and $\hat{\beta} \pm 1.96\text{se}\hat{\beta}$. But β is fixed, if unknown. It is not stochastic! What we are actually saying is that there is a 95% chance that the confidence interval we construct using our estimates, that is, $\hat{\beta} \pm 1.96\text{se}\hat{\beta}$, in fact contains the true value β .

What *is* important is the fact that we should be using the t -distribution (Section II.E.4.6) rather than the normal distribution because we have to estimate these standard errors. For every parameter whose value is unknown and whose value therefore has to be estimated, we lose one ‘degree of freedom’ when manipulating the data. In a sense, the estimation of the parameter acts like a constraint that has to be satisfied. Hence, the number of degrees of freedom in the t -distribution is the number of data points less the number of parameters that are estimated in the model, that is, $n - k$.

To illustrate this we construct a 95% confidence interval for β from the output of the simple regression in Section II.F.1.3. In that regression the point estimate of β is 1.265945 and the estimated standard error is 0.127829 (both to 6 decimal places). Using Excel we find that entering ‘=TINV(0.05,61)’ into a cell returns the value 1.999624. (The 0.05 refers to the combined tail probabilities of the t -distribution, and we have 61 degrees of freedom – 63 data items less two parameters estimated from the data.) Thus the 95% confidence interval is from $1.265945 - 1.999624 \times 0.127829$ to $1.265945 + 1.999624 \times 0.127829$, that is, from 1.0103 to 1.5216 (to 4 decimal places). This is often written as (1.0103, 1.5216).

Using the critical value 1.999624 from the t -distribution in equation (II.F.14) instead of 1.959963 from the normal distribution produces a slightly wider confidence interval. This reflects slightly less confidence in our point estimates than would be the case if we knew the correct values for the measures of variability. Similar reasoning produces a 95% confidence interval of (–0.0174, 0.0157) for α . Note that this interval includes zero. We can therefore not exclude zero as a possible value for α at the 95% level of confidence.

II.F.5 Hypothesis Testing

A *statistical hypothesis* is an assumption about the value of a population parameter of the probability distribution under consideration. Two hypotheses are established, the *null hypothesis* and the *alternative hypothesis*. In effect we set up two competing hypotheses and test which of the two applies.

The null hypothesis, usually designated H_0 , is the assumption that will be accepted pending evidence against it. The alternative hypothesis, usually designated H_1 , is that hypothesis which will be accepted if the statistical test leads to a rejection of the null hypothesis.

The exact formulation of the hypothesis depends upon what we are trying to establish. For example, imagine that we simply wish to know whether or not a population parameter, say the

mean, μ , has a value of μ_0 . The hypotheses would then be formulated as

$$H_0: \mu = \mu_0, \quad H_1: \mu \neq \mu_0.$$

However, if we wished to know whether or not a population parameter is greater than a given value, μ_0 , the hypothesis (in relation to means) would be

$$H_0: \mu = \mu_0, \quad H_1: \mu > \mu_0.$$

If we wished to know whether the population parameter was less than μ_0 , the hypotheses would be

$$H_0: \mu = \mu_0, \quad H_1: \mu < \mu_0.$$

The first of these is known as a *two-tailed* test. The other two are both *one-tailed* tests.

A statistical test consists of using a statistic, computed from a sample, to decide whether or not to reject a hypothesis about the population parameter in question. The procedure is as follows:

1. Decide on the level of significance for the test. This will usually be 10%, 5% or 1%. For a 5% one-tailed test we need to arrange for there to be a tail probability of 5%. For a two-tailed test the tail probabilities are each set to 2.5%.
2. Set up the null and alternative hypotheses.
3. Choose an appropriate statistic for the test – this is called the ‘test statistic’.
4. Identify the appropriate critical values – those values of the test statistic which lead to the rejection of the null hypothesis.
5. Use the data to find a value for the test statistic.
6. Apply the decision rule: Accept H_0 if the computed value of the statistic is not in the critical region. Reject H_0 and accept H_1 if the computed value of the statistic is in the critical region.

An alternative but equivalent view of hypothesis testing is to find the *probability value* (or *p-value*) of the test statistic’s value. This is a probability – the probability of getting a value for the test statistic that is equal to or more extreme than the value observed, assuming that the null hypothesis is true.

II.F.5.1 Significance Tests for the Regression Parameters

Again we refer to the simple regression of Section II.F.1. Following the procedure just outlined, we will test $H_0: \beta = 0$ against $H_1: \beta > 0$ at the 2.5% level of significance.¹² For a variable that has a

¹² The Excel command we use gives a two-tailed value, so asking for 5% is equivalent to asking for a one-tailed value at 2.5%.

normal distribution, the estimated mean divided by the estimate of its standard deviation has a t -distribution. Our estimate has 61 degrees of freedom (63 data items less two parameters estimated from the data), so our test statistic will have a t -distribution with 61 degrees of freedom. The Excel command ‘TINV(0.05, 61)’ returns the value 1.999624. So our critical region consists of values of the test statistic which exceed 1.999624. The computed value of the test statistic is

$$\frac{\text{the estimate of } \beta}{\text{the estimate of } \text{se}\beta} = \frac{1.26594499}{0.12782945} = 9.903390767.$$

This is in the critical region, so we reject H_0 and accept H_1 .

The reader will note the similarities between this procedure and the construction of a confidence interval in Section II.F.4.1. Alternatively, we could have found the p -value for our test statistic from the Excel command ‘=TDIST(9.9034, 61, 1)’. This returns the value 1.29063E-14, this being shorthand for 1.29063×10^{-14} or 0.0000000000000129063. It shows that if β had the value zero then our observation would have been very unlikely indeed. This information is shown on the Excel output, except that the displayed p -values are for two tails. Thus the information displayed includes a p -value of 2.58×10^{-14} for the slope estimate. For the intercept it can be seen that the value of the t -statistic is -0.1018 , with a p -value of 0.9192 (two-tailed), or 0.4596 (one-tailed). This is not at all unlikely – we cannot reject the possibility that the true value of α is zero.

II.F.5.2 Significance Test for R^2

The computed value of R^2 is also an estimate. Its significance is tested using the F -distribution. The F -distribution has two sets of degrees of freedom, one (often designated ν_1) in the numerator of the test statistic and a second (designated ν_2) in the denominator. In our simple regression in Section II.F.1.3 there was one degree of freedom in ESS (being the number of explanatory variables) and 61 degrees of freedom in RSS (being the number of observations less the number of explanatory variables). The ratio of the quantities divided by their respective degrees of freedom is F -distributed, that is,

$$\frac{\text{RSS}/1}{\text{ESS}/61} \sim F_{61}^1.$$

But

$$\frac{\text{RSS}/1}{\text{ESS}/61} = \frac{61}{1} \times \frac{\text{RSS}}{\text{ESS}} = \frac{61}{1} \times \frac{\text{RSS}}{\text{TSS} - \text{RSS}} = \frac{61}{1} \times \frac{R^2}{1 - R^2} \quad \text{since } R^2 = \frac{\text{RSS}}{\text{TSS}}.$$

So our test statistic takes the value

$$\frac{61}{1} \times \frac{0.616538261}{1 - 0.616538261} \cong 98.08.$$

Using Excel we find that entering ‘=FDIST(98.08,1,61)’ into a cell returns the value 2.57793E-14, or 2.57793×10^{-14} . It is very small. It represents the probability of the test statistic being 98.08 or larger if, in fact, the true value of R^2 was zero. We thus reject that possibility – the true value of R^2 is not zero and the model appears to have explanatory value. By itself, this is neither surprising nor interesting. However, the methodology is developed in multiple regression to provide the means for testing whether or not introducing or deleting a variable makes a significant difference to the model.

II.F.5.3 Type I and Type II Errors

When testing hypotheses there is a possibility that the null hypothesis will be rejected when in fact it should have been accepted. This is referred to as a *type I error*. The probability of making a type I error is the significance level of the test. Thus when we choose a 5% level of significance for our test, we are accepting a 5% chance that we will reject the null hypothesis when in fact we should have accepted it. The second type of error is that the null hypothesis is accepted when in fact it should have been rejected. This is referred to as a *type II error*.

Type I and type II errors are possible whenever a decision has to be made. Consider, for example, the problem facing a jury in England. The defendant is either innocent or guilty. In the English judicial system the jury can find the defendant innocent or guilty. The jury system is actually testing the null hypothesis that the defendant is not guilty. If the jury finds the defendant guilty when he or she is indeed guilty that is good – there is no error. Likewise, if the jury finds the defendant innocent when he or she is indeed innocent all is well and good. But what if the defendant is innocent but the jury finds the defendant guilty, that is, they reject the null hypothesis when actually it is correct? This is a type I error. It is the type of error that one most wants to avoid. Accordingly, we set the significance level of the test to make the probability of this small (commonly 5% or 1%). Having set the significance level we have little influence over the probability of a type II error – in this case, finding a guilty person to be innocent.

II.F.6 Prediction

A regression model will naturally be used to explore possible outcomes for values of the independent variables which have not been observed. This is often referred to as *prediction*. There are two associated questions:

1. What is the probability distribution of \hat{Y} ?
2. What is the probability distribution of Y ?

Both are normal. Both are centred on $E(Y)$, which we estimate by our computed value of \hat{Y} . But they differ in their variabilities, as measured by their standard errors. In the case of simple linear regression \hat{Y} represents the height of the regression line at the chosen value of X . The line goes through $(0, \hat{\alpha})$ and has a gradient of $\hat{\beta}$, so one can visualise how the height of the line at the chosen value of x depends on the sampled values of α and β . The dependency on β is particularly important if the chosen value of X is outside of the range of the observed X s, that is to say, if we are *extrapolating*. But the height of the line only represents the expected value of Y at the chosen value of X . There is then the variation about the central value to take into account when considering the probability distribution for a single realisation. In summary, Y has more variability than $E(Y)$, because Y has the additional uncertainty of the error variable, ϵ .

Most statistical packages provide both of these standard errors, the standard error of the prediction mean and the standard error of predictions. Otherwise there are formulae from which they can be computed. That for the standard error of the prediction interval is:

$$s \sqrt{1 + \frac{1}{n} + \frac{(X^* - \bar{X})^2}{\sum (X_i - \bar{X})^2}} \quad (\text{II.F.15})$$

where \bar{x} is the chosen value of x and where

$$s = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n - 2}}.$$

The formula for the standard error of \hat{Y} is the same, but without the ‘1’, that is,

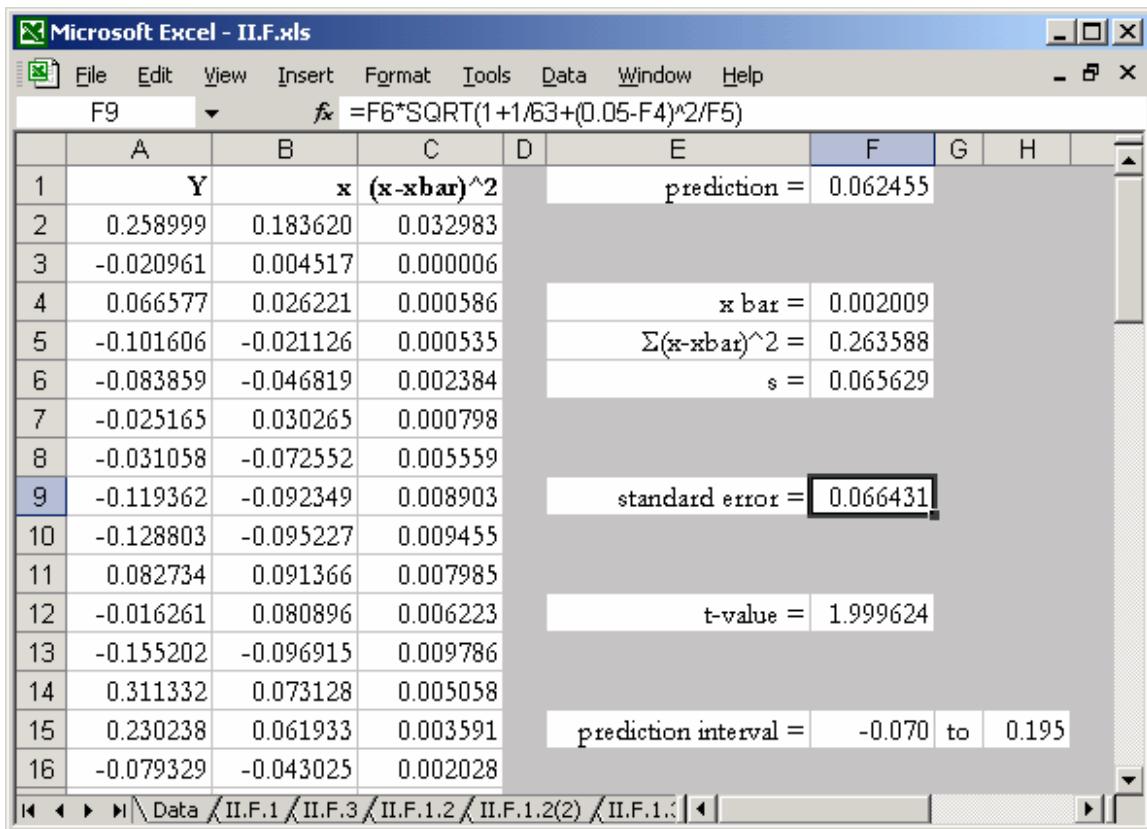
$$s \sqrt{\frac{1}{n} + \frac{(X^* - \bar{X})^2}{\sum (X_i - \bar{X})^2}}.$$

(The difference between the two variances (given by squaring the expressions) is s^2 , which represents the variability about the regression line.)

Excel is not a statistical package as such. Its ‘=FORECAST’ function makes the computation of \hat{Y} easy for any given value of x , but the standard error of \hat{Y} and of Y are not available, save by implementing the formulae. (But note that s is available – it is the square root of (the residual sum of squares divided by its number of degrees of freedom) – 0.065628685 in the simple linear regression example in Section II.F.1.3. It is displayed in the output as ‘Standard Error’.)

To demonstrate the computation we calculate a 95% confidence interval for the returns to the Hilton equity when the FTSE 350 L&H index return takes a value of 0.05 (see Figure II.F.7). Therefore we can be 95% confident that if the index return is 0.05, then the return on Hilton Hotels will lie between -0.070 and 0.195, with the expected value being 0.062.

Figure II.F.7: [Predicting the return on Hilton Hotels shares](#)



II.F.7 Breakdown of the OLS Assumptions

The inferential part of an OLS regression analysis (i.e. constructing confidence intervals and/or performing tests of significance) is based upon a number of assumptions. The validity of the hypothesis testing and confidence interval construction is dependent on those assumptions being true. Therefore, it is important to determine whether the assumptions of OLS regression have been satisfied. We can test for:

- *heteroscedasticity*, which is when the residuals do not have a constant variance;
- *autocorrelation*, which exists if the residuals are not independent;
- *multicollinearity*, which occurs in multiple regression applied to time series when some or all of the independent variables are correlated to some degree.

II.F.7.1 Heteroscedasticity

If the residuals have a constant variance they are said to be homoscedastic, but if they are not constant they are said to be heteroscedastic. The consequence of heteroscedasticity on the prediction interval estimation and hypothesis testing is that although the coefficients are unbiased, the variances, and therefore the standard errors, of those coefficients will be biased. If this bias is negative, the estimated standard errors will be smaller than they should be and the test statistic will be larger than it is in reality. Thus we may be led to think that the coefficient is significant when it is not. Conversely, if the bias is positive, the estimated standard errors will be larger than they should be and the test statistics smaller. We may therefore accept the null hypothesis when in fact it should be rejected. There are many ways of attempting to deal with heteroscedasticity once it is identified. However, this is beyond the scope of the PRM syllabus.

II.F.7.2 Autocorrelation

Autocorrelation, also known as serial correlation, occurs in time-series data when the errors are not independent of each other because current values of Y are influenced by past values. The errors are then autocorrelated. For example, assume that the error ε_t is related to the error in the previous time period ε_{t-1} . Then it might be described by an autoregressive relationship of the form

$$\varepsilon_t = \gamma\varepsilon_{t-1} + \varepsilon_t^*$$

where ε_t^* is another random variable. The right-hand side of this expression is referred to as a first-order autoregressive function, denoted by 'AR(1)'. Only one preceding time period is incorporated into the function. A second-order, AR(2), model might take the form:

$$\varepsilon_t = \gamma_1\varepsilon_{t-1} + \gamma_2\varepsilon_{t-2} + \varepsilon_t^*$$

The OLS regression model is a minimum variance, unbiased estimator only when the residuals are independent of each other. If autocorrelation exists in the residuals, the regression coefficients are unbiased but the standard errors will be underestimated and the tests of regression coefficients will be unreliable. There are many tests for autocorrelation but, again, they are beyond the scope of the PRM syllabus.

Autocorrelation in the errors may be caused by omitted variables or by using the wrong functional form of the estimating equation – for example, using a linear model when it should be non-linear. In this case, the way to deal with it is to include more, relevant, variables as explanatory variables. If that is impossible, special estimation algorithms are available (such as the ‘Cochrane–Orcutt’ procedure) but again, we do not deal with this in the PRM course. Introducing lagged variables can also cause autocorrelation – often negative autocorrelation. This is when large positive values tend to be followed by large *negative* values, and conversely. However, positive autocorrelation is much more common than negative correlation.

II.F.7.3 Multicollinearity

When some or all of the independent variables in a multiple regression are correlated, it is difficult or impossible to untangle their separate explanatory effects on Y .

For instance, suppose that $Y = 5X_1 + 2$, but that $X_1 = 2X_2$. Then by substitution we can see that we can just as well express Y as $10X_2 + 2$. Equally, we could have chosen to substitute, say, just 2 of the X_1 s, giving $Y = 3X_1 + 4X_2 + 2$. There are myriad expressions for Y , all of them equivalent. This example corresponds to a situation in which the correlation between X_1 and X_2 is 1, but similar effects are seen in stochastic situations when the independent variables are jointly correlated. This is known as *multicollinearity*.

There is no exact boundary value of the degree of correlation between variables that causes this problem of multicollinearity, and judgement is required in dealing with it. It is particularly prevalent in multi-factor models that use time-series data to model equity returns. Style factors, industry indices and even major market factors can be highly correlated. For this reason some equity analysis companies employ principal components analysis (PCA) on the factors (see Section III.A.3.7). Since principal components are uncorrelated by definition, a regression on the principal components of the factors, instead of the factors themselves, is a common way of avoiding multicollinearity.

With multicollinearity the regression coefficients are unstable in the degree of statistical significance, magnitude and sign. The R^2 may be high but the standard errors are also high and, consequently, the t -statistics are small, indicating apparent lack of significance.

Apart from using PCA, multicollinearity can be remedied in several other ways. However each of the following methods has problems:

1. *Add further sample data.* This might be considered on the basis that correlation between factors may not be so high in an extended sample. This is not always possible and, even when possible, it often does not work!
2. *Drop those variables that are highly correlated with the others.* The problem here is that presumably the variables were included on theoretical grounds and it is inappropriate to exclude them just to make the statistical results ‘better’.
3. *Pooling of cross-section and time-series data.* This technique takes a coefficient from, say, a cross-section regression and substitutes it for the coefficient of the time-series equivalent data. This is not always possible and, even when possible, it makes the strong assumption that the cross-sectional relationship is identical to the time-series relationship between the variables.

II.F.8 Random Walks and Mean Reversion

In the statistical analysis of financial time-series data there is a fundamental distinction to be drawn between returns data (which are usually ‘mean-reverting’ or ‘stationary series’) and price data (which are normally ‘random walks’ or, at least, ‘integrated series’). This section explains how regression analysis may be used to distinguish between these two types of time series.

A formal definition of *stationarity* includes the property that the (unconditional) mean and the variance of the time series are finite constants. In this case, the series will be ‘mean-reverting’ because, with a finite variance, it can never drift too far from its mean. The calculation of correlation, and of volatility (or standard deviation), always assumes that the series concerned is stationary.

Never try to estimate a correlation or a volatility using raw price data! Prices are usually not a stationary series; in fact they are often random walks. A random walk is a stochastic process for a *level* variable whose increments are determined by a zero-mean, i.i.d. stationary random variable. For instance, the level variable of the random walk might be the log of asset prices, in which case the stationary increment will be the asset returns.

A random walk model, with *level* variable Y_t and *drift* α is specified as:

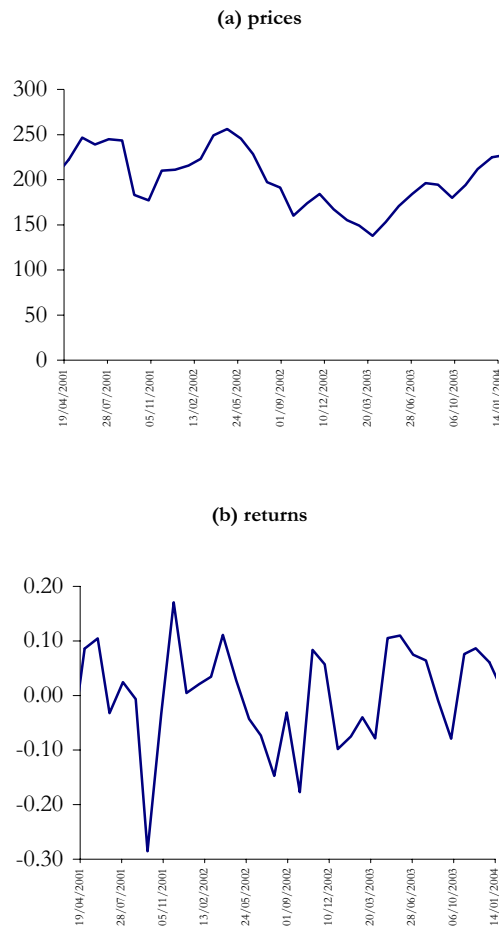
$$Y_t = \alpha + Y_{t-1} + \varepsilon_t$$

where the *increment* ε_t is i.i.d. If $\alpha > 0$ the random walk has an upward drift; if $\alpha < 0$ the random walk has a downward drift. A random walk is a particular type of *integrated* time series. The term ‘integration’ refers to the degree of differencing that a data series requires for it to be transformed into a stationary series. Differencing is the process of finding the change in the value of a variable in successive time periods, e.g. $\Delta Y_t = Y_t - Y_{t-1}$. The ΔY series is the differenced series.

If a time series has to be differenced only once in order to be transformed into a stationary series, the original series is said to be integrated of order 1, or $I(1)$. If the original series has to be differenced twice to become stationary, the original series is said to be integrated of order 2, or $I(2)$. If a series does not have to be differenced at all because it is already stationary, it is said to be integrated of order 0, or $I(0)$.

Intuitively, we would not expect many time series of share prices, exchange rates or index levels to be stationary because rising or falling values over time are a feature of financial variables. However, the returns required by investors should be dependent only on the uncertainty surrounding the investment, and should be independent of the asset price or index level. Thus returns data may have a constant mean and standard deviation.

Figure II.F.8: (a) Prices and (b) returns of Hilton shares



To illustrate these points consider Figures I.F.8. Figure I.F.8(a) shows the prices of the Hilton shares from January 2001 to June 2004. At the beginning of the period the drift does not appear to be significantly positive or negative. However, after a time there are long periods of trends in these prices: between May 2002 and May 2003 the drift in prices appears negative, after that it appears positive. Now look at Figure I.F.8(b). This shows the returns data for the same time period as Figure I.F.8(a). The returns data oscillate around a constant mean; there is no drift in the returns.

The efficient markets hypothesis¹³ states that security prices embody all available information and hence using resources to identify under- or overpriced securities is futile. A test of the so-called ‘weak form’ market efficiency is a test to determine whether asset prices follow a random walk.

¹³ There is a huge literature on this subject; interested readers should study the papers by Fama (1970) and Malkiel (2003).

Given a time series Y_t , we can test whether it is integrated, with the alternative that it is a stationary series, by regressing ΔY_t against Y_t :

$$\Delta Y_t = \alpha_1 + \alpha_2 Y_{t-1} + \varepsilon_t.$$

This is called a ‘Dickey–Fuller’ regression. If α_2 is not significantly different from zero, then Y is integrated of order 1 (so ΔY will be stationary) but if α_2 is significantly *less* than zero, then Y is already stationary. Note that the statistic used to test whether $\alpha_2 = 0$ is not an ordinary ‘ t -statistic’ as given in Section II.F.5.1. The Dickey–Fuller statistic is, like the t -statistic, the ratio of the estimated coefficient to its estimated standard error, but its critical values are higher than the usual t -distributed critical values (see Dickey and Fuller, 1979).

Applying a Dickey–Fuller regression to the Hilton prices produces a value for the Dickey–Fuller statistic of -2.47 , which is not significant.¹⁴ So we cannot reject the hypothesis that the price series is integrated. However the Dickey–Fuller test statistic for the returns data has the value -7.01 , so we can reject the hypothesis that the series is integrated and accept the alternative hypothesis that it is stationary.

II.F.9 Maximum Likelihood Estimation

In Section II.F.1 we discussed parameter estimation in a regression model using the OLS method. Maximum likelihood estimation (MLE) is an alternative method of estimating parameters in a regression model. In fact, MLE is a more general approach to estimating parameters in statistical models as it can be used in many circumstances where OLS would be unsuitable.

For MLE we must make an assumption about the functional form of the distribution that generates the data. Given the functional form specified, one determines the *likelihood function*, L , which is the probability distribution of the entire sample. This function will depend on the unknown parameters (i.e. α and β in the case of a simple regression model). The MLE parameter estimates will be values chosen to maximise the likelihood function, or equivalently the logarithm of the likelihood function, since the parameter value that maximizes L are the same as the value that maximizes $\ln(L)$. Normally, the log-likelihood function is more convenient to work with than the likelihood function itself. Maximisation can be achieved using either analytical or numerical methods, as will be explained in Chapter II.G. The analytic approach is to differentiate the log-likelihood function. For complex models with multiple parameters numerical methods are often necessary.

¹⁴ The critical value of the Dickey–Fuller statistic depends on whether a constant, trend and/or a lagged dependent variable are included in the regression, as well as on the significance level of the test. It is tabulated in standard time series texts such as Hamilton (1994) and is commonly in excess of -3 .

For illustration, we can apply this approach to the classical regression model, with the assumption that the error terms are i.i.d. normal. We differentiate the log-likelihood function with respect to β and set the derivative equal to zero. Solving this expression, we obtain an estimate of β which is identical to the estimate obtained using the method of OLS. That is, under the standard assumptions the two methods are equivalent. If the standard assumptions do not hold, MLE can still be applied. The resultant parameter estimates will be *consistent* estimators. This means that, although the estimators may be biased in small samples, they will approach their true values for large samples.

II.F.10 Summary

This chapter is about linear regression models – models used to analyse the relationship between two or more variables. The main method for estimating the parameters is ordinary least squares, although an alternative, maximum likelihood estimation, is also briefly considered. The method of ordinary least squares is described both for simple and multiple regression models. The chapter also explains how to interpret the estimation results and determine whether they are statistically significant using significance tests. Some of the major problems encountered in ordinary least-squares analysis are briefly discussed, along with some suggested solutions.

References

- Dickey, D A, and Fuller W A (1979) Distribution of the estimator for autoregressive time series with a unit root, *Journal of the American Statistical Association*, 74, pp. 427–431.
- Fama, E (1970) Efficient capital markets: a review of theory and empirical work, *Journal of Finance*, 25, pp. 383–417.
- Hamilton, J D (1994) *Time Series Analysis*. Princeton, NJ: Princeton University Press.
- Malkiel, B G (2003) The efficient markets hypothesis and its critics, *Journal of Economic Perspectives*, 17(1), pp. 59–82.
- Ross, S (1976) The arbitrage theory of capital pricing, *Journal of Economic Theory*, 13, pp. 341–360.
- Sharpe, W F (1964) Capital asset prices: A theory of market equilibrium under conditions of risk, *Journal of Finance*, 19, pp. 425–442.
- Watsham, T J, and Parramore, K (1997) *Quantitative Methods in Finance*. London: Thomson Learning.

II.G Numerical Methods

Keith Parramore and Terry Watsham¹

Numerical methods are used when analytic (or exact, or ‘closed form’) solutions are too complex or just not available. This chapter explains the mathematical basis for numerical methods and discusses the applications of some common numerical techniques in mathematical finance. We look at methods for solving ordinary equations, of which Excel’s ‘Goal Seek’ is an implementation. We look at optimisation. Excel’s optimiser is called ‘Solver’ – which is something of a misnomer since Goal Seek is the equation solver! We also look at the use of finite differences and binomial lattices for valuing options, and revisit the use of simulation, showing how to use it to value an Asian option.

By the end of this chapter you will:

- be able to formulate as equations problems involving finding the yield (internal rate of return) from a number of cash flows and payments, finding loan interest rates (APRs) given repayment schedules, and finding the implied volatility of an asset given the price of an option on it;
- be able to solve such equations by iterative methods, including Goal Seek;
- have an appreciation of the mathematical methods underpinning equation solvers such as Goal Seek;
- understand the essential features of an optimisation problem, and know the difference between unconstrained optimisation and constrained optimisation;
- be able to use Solver on unconstrained and constrained optimisation problems;
- have an outline knowledge of how optimisers such as Solver work, and an appreciation of their limitations;
- know how to value American options using binomial lattice methods and finite difference methods, and how to value Asians using simulation.

¹ Keith Parramore is Principal Lecturer in Mathematics at the University of Brighton and Terry Watsham is Principal Lecturer at the University of Brighton.

II.G.1 Solving (Non-differential) Equations

II.G.1.1 Three Problems

We first set up three problems – each easy to formulate but not so easy to solve.

1. An investment opportunity will return 5000 at the end of year 1, 4000 at the end of year 2, 4000 at the end of year 3, 2500 at the end of year 4 and 53,000 at the end of year 5, the latter including the realisation of assets. The investment will cost 50,000. What is the yield?
2. A loan of 50,000 is to be repaid by 25 equal annual payments of 3500. What is the interest rate?
3. A European option is currently trading at 4.42. The option is to purchase for 105 in six months' time an asset currently valued at 100. The risk-free rate of interest is 4%. What is the implied volatility of the asset?

The first two of these problems can both be formulated as polynomial equations, the first of degree 5 and the second of degree 26. The third is more difficult to specify explicitly since it involves two separate evaluations of the cumulative normal distribution function.

1. Let the yield be $y\%$ and put $x = \left(1 + \frac{y}{100}\right)$. If we can solve this to find x , then we will

know y . We have $50,000 = \frac{5000}{x} + \frac{4000}{x^2} + \frac{4000}{x^3} + \frac{2500}{x^4} + \frac{53,000}{x^5}$. This simplifies to:

$$50,000x^5 - 5000x^4 - 4000x^3 - 4000x^2 - 2500x - 53,000 = 0. \quad (\text{II.G.1})$$

2. Let the annual compounding factor be x , that is, $x = 1 + y$. If we can find x then we will know the interest rate y . Throughout the first year of the loan the debt is 50,000. At the start of the second year the debt is $50,000x - 3500$. Similarly, the debt at the start of the third year is $(50,000x - 3500)x - 3500$. Following the same approach, the debt at the start of the 26th year is

$$(\dots(((50,000x - 3500)x - 3500)x - 3500)\dots)x - 3500,$$

where the ‘...’ indicates 21 more applications of multiplying by x and subtracting 3500.

This expands to $50,000x^{25} - 3500x^{24} - 3500x^{23} - 3500x^{22} - \dots - 3500$. Using the formula for the sum of a geometric progression (see Section II.A.2.2)² it simplifies to

$$50,000x^{25} - 3500\left(\frac{x^{25} - 1}{x - 1}\right).$$

² If $S_n = a + ar + ar^2 + ar^3 + \dots + ar^{n-1}$, then S_n is called ‘the sum to n terms of a geometric progression’. By multiplying through by r , and looking at the difference between S_n and rS_n , it can be shown that $S_n = (1 - r^n)/(1 - r)$.

Since the debt must be paid off by the start of the 26th year, we can set the expression above equal to zero. After simplifying that gives³

$$50,000x^{26} - 53,500x^{25} + 3500 = 0. \quad (\text{II.G.2})$$

3. The Black–Scholes formula described in Section I.A.8.7 gives the value of a European call or a European put. In its simplest case the formula is:

$$V = SN(d_1) - Xe^{-r(T-t)}N(d_2) \quad (\text{II.G.3a})$$

for a European call, and

$$V = -SN(-d_1) + Xe^{-r(T-t)}N(-d_2) \quad (\text{II.G.3b})$$

for a European put.

Here

$$d_1 = \frac{\ln\left(\frac{S}{X}\right) + \left(r + \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}},$$

$$d_2 = \frac{\ln\left(\frac{S}{X}\right) + \left(r - \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}$$

and t is the time now, T the expiry time, S the present price of the security, X the strike price, σ the volatility, r the risk-free rate of interest, and N the cumulative normal distribution function.

Having formulated these problems we will now look at alternative approaches to solving them.

II.G.1.2 Bisection

In a mathematical parlour trick a guest is invited to select, but not reveal, a word from a dictionary. The mathematician then selects a word from (roughly) halfway through the dictionary and asks the guest whether the selected word is before or after it alphabetically. The mathematician then repeats the process using the first or second half of the dictionary. At each stage the set of possible words is halved. In a dictionary containing 100,000 words the mathematician will certainly need to ask no more than 17 questions to home in on the chosen

word, since $2^{17} > 100,000$, so $\frac{100,000}{2^{17}} < 1$.

³ Note that $x = 1$ is a solution to this equation. That is a consequence of multiplying both sides by the factor $(x - 1)$ during the simplification. However, a solution to our problem is a value of x which satisfies equation (II.G.2) and which is larger than 1.

Example II.G.1

Using the same approach to problem 1, equation (II.G.1), we know the solution lies between 1 and 2. Let $f(x) = 50,000x^5 - 5000x^4 - 4000x^3 - 4000x^2 - 2500x - 53,000$. Now $f(1) = -18,500$ and $f(2) = 1,414,000$, and we consider $x = 1.5$. We have $f(1.5) = 275,125$, so the answer must lie between 1 and 1.5 since we are searching for x such that $f(x) = 0$.

Consider $x = 1.25$. $f(1.25) = 70,193.26$, so the answer lies between 1 and 1.25.

Consider $x = 1.125$. $f(1.125) = 15,522.28$, so the answer lies between 1 and 1.125.

Consider $x = 1.0625$. $f(1.0625) = -3,637.82$, so the answer lies between 1.0625 and 1.125.

Figure II.G.1: Bisection with Excel (1)

	Left	Midpoint	Right	f(midpoint)
3	1	1.5	2	275125
4	1	1.25	1.5	70193.35938
5	1	1.125	1.25	15522.27783
6	1	1.0625	1.125	-3637.815475
7	1.0625	1.09375	1.125	5354.859054
8	1.0625	1.078125	1.09375	718.1805614
9	1.0625	1.0703125	1.078125	-1494.110886
10	1.0703125	1.0742188	1.078125	-396.6365824
11	1.0742188	1.0761719	1.078125	158.5917962
12	1.0742188	1.0751953	1.0761719	-119.5658956
13	1.0751953	1.0756836	1.0761719	19.3768817
14	1.0751953	1.0754395	1.0756836	-50.12849994
15	1.0754395	1.0755615	1.0756836	-15.38431039
16	1.0755615	1.0756226	1.0756836	1.994159964
17	1.0755615	1.075592	1.0756226	-6.695606589
18	1.075592	1.0756073	1.0756226	-2.350856163
19	1.0756073	1.0756149	1.0756226	-0.178381313

At each stage we evaluate at the midpoint and choose the half interval which traps the solution, that is, the half interval which gives opposite signs for $f(x)$ at its extremities. We repeat this process until the size of the interval, the interval of uncertainty, is as small as we please. At each stage our guess is at the midpoint of the interval, and our *error bound* is half the width of the interval.

The key to implementing this in Excel is to use $f(\text{midpoint})$ on one row to decide what are the 'left' and 'right' values on the next row. The '=IF' function is used for this, and this is shown in Figure II.G.1. Once the second row is coded (row 4 of the spreadsheet) it can be copied down as far as is required. We can see from row 18 that the yield is 7.56%.

Example II.G.2

Applying the method of bisection to problem 2, we first note that the solution also lies between 1 and 2, as shown in Figure II.G.2. The only difference between this and Example II.G.1 is in the formula in column E. In row 19 the interest rate can be seen to be either 4.86% or 4.87% to 2 decimal places – more iterations would be needed to determine which.

Example II.G.3

For problem 3 we set up a user-defined function, BSC (price, strike, volatility, risk-free rate, time) for evaluating the Black–Scholes value (see Figure II.G.3). All the inputs are known or user-defined, with the exception of volatility – the value we wish to infer from the option's market value. We assume initially that the volatility is between 1% and 50%, (see cells B3 and D3). Using the bisection method we hone in on the value of volatility which will equate the Black–Scholes model price to the observed market price of \$4.42 (equivalently, the Black–Scholes price less 4.42 equal to zero). From the results displayed in Figure II.G.3 we can be confident that the implied volatility is 20.15% to 4 significant figures.

Figure II.G.2 [Bisection with Excel \(2\)](#)

	Left	Midpoint	Right	f(midpoint)
3	1	1.5	2	542903618.3
4	1	1.25	1.5	2385780.164
5	1	1.125	1.25	55757.15415
6	1	1.0625	1.125	1792.916613
7	1	1.03125	1.0625	-681.5832757
8	1.03125	1.046875	1.0625	-134.3141389
9	1.046875	1.0546875	1.0625	601.9152791
10	1.046875	1.0507813	1.0546875	184.8513326
11	1.046875	1.0488281	1.0507813	13.91924226
12	1.046875	1.0478516	1.0488281	-62.92951688
13	1.0478516	1.0483398	1.0488281	-25.20106635
14	1.0483398	1.048584	1.0488281	-5.816529632
15	1.048584	1.0487061	1.0488281	4.007246154
16	1.048584	1.048645	1.0487061	-0.915643498
17	1.048645	1.0486755	1.0487061	1.543047669
18	1.048645	1.0486603	1.0486755	0.313014073
19	1.048645	1.0486526	1.0486603	-0.301486665

Figure II.G.3 [Bisection with Excel \(3\)](#)

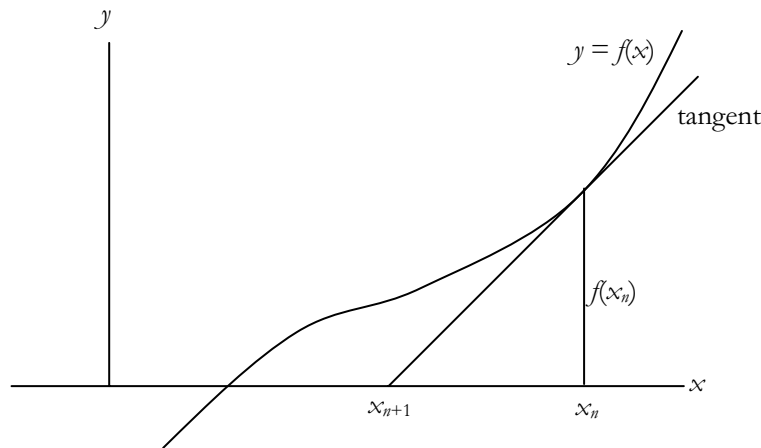
	Left	Midpoint	Right	f(midpoint)
3	0.01	0.255	0.5	1.50060621
4	0.01	0.1325	0.255	-1.912034241
5	0.1325	0.19375	0.255	-0.217666215
6	0.19375	0.224375	0.255	0.639871111
7	0.19375	0.2090625	0.224375	0.210589292
8	0.19375	0.2014063	0.2090625	-0.003684435
9	0.2014063	0.2052344	0.2090625	0.10341834
10	0.2014063	0.2033203	0.2052344	0.049858143
11	0.2014063	0.2023633	0.2033203	0.023084615
12	0.2014063	0.2018848	0.2023633	0.009699525
13	0.2014063	0.2016455	0.2018848	0.003007403
14	0.2014063	0.2015259	0.2016455	-0.000338551
15	0.2015259	0.2015857	0.2016455	0.001334417
16	0.2015259	0.2015558	0.2015857	0.000497931
17	0.2015259	0.2015408	0.2015558	7.96893E-05
18	0.2015259	0.2015334	0.2015408	-0.000129431
19	0.2015334	0.2015371	0.2015408	-2.4871E-05

II.G.1.3 Newton–Raphson

Bisection has the advantage of security of knowledge – we always have bounds on where the solution lies. But we need to start with appropriate bounds in the first place, and that is a disadvantage with respect to automation. An alternative approach is *formula iteration* where a guess is repeatedly replaced by $f(\text{guess})$ for some function f . There are often many appropriate formulae for a given equation. The Newton–Raphson formula almost always works, and has fast convergence.

The method uses the tangent to the function curve ($y = f(x)$) at the point given by the current guess (x_n).

Figure II.G.4: Newton–Raphson



From Figure II.G.4 we can see that the gradient of the tangent is

$$f'(x_n) = \frac{f(x_n)}{x_n - x_{n+1}}.$$

This solves to give the Newton–Raphson formula:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{II.G.4}$$

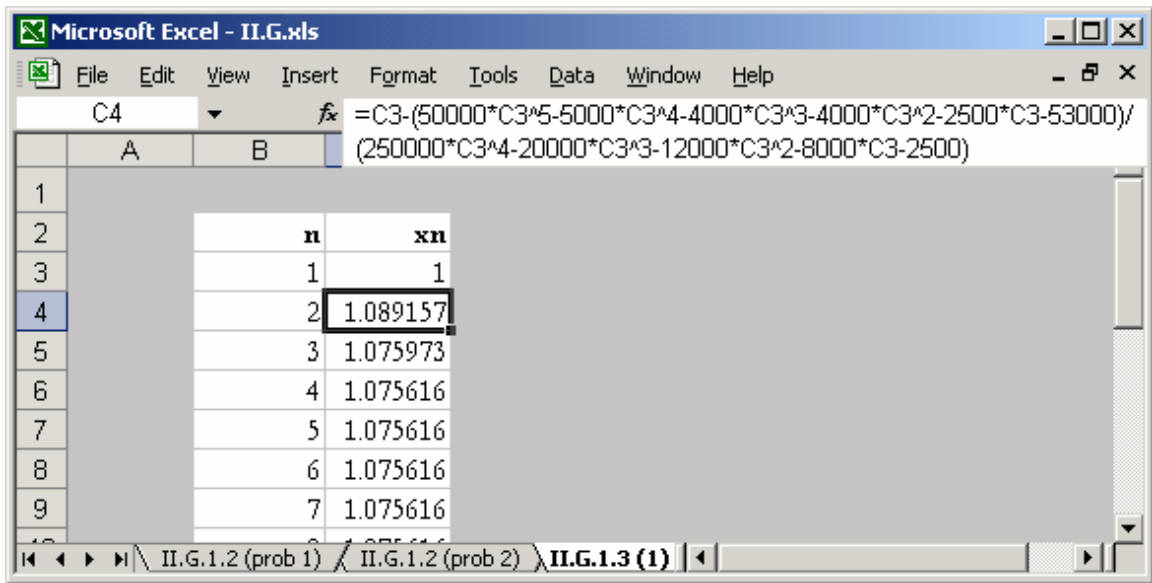
Example II.G.4: Problem 1

$$f(x) = 50,000x^5 - 5000x^4 - 4000x^3 - 4000x^2 - 2500x - 53,000$$

$$f'(x) = 250,000x^4 - 20,000x^3 - 12,000x^2 - 8000x - 2500.$$

Implementing this in Excel, we can see that convergence to 7.56% is very rapid indeed (see row 5 of Figure II.G.5).

Figure II.G.5 [Newton–Raphson using Excel \(1\)](#)



Example II.G.5: Problem 2

$$f(x) = 50,000x^{26} - 53,500x^{25} + 3500$$

$$f'(x) = 1,300,000x^{25} - 1,337,500x^{24}$$

Figure II.G.6: [Newton–Raphson using Excel \(2\)](#)

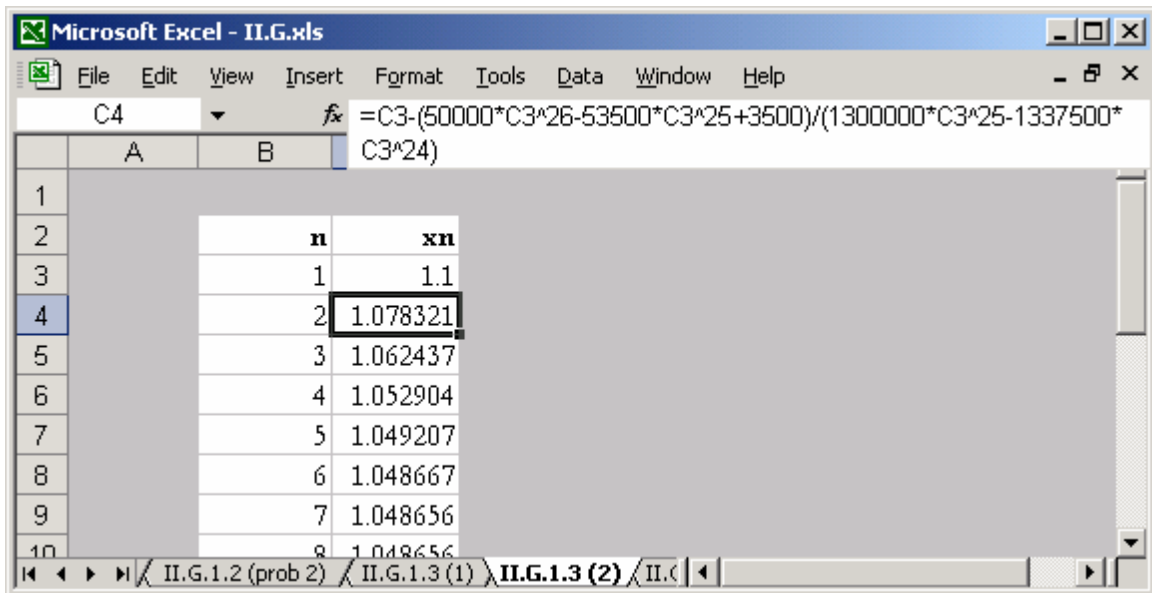


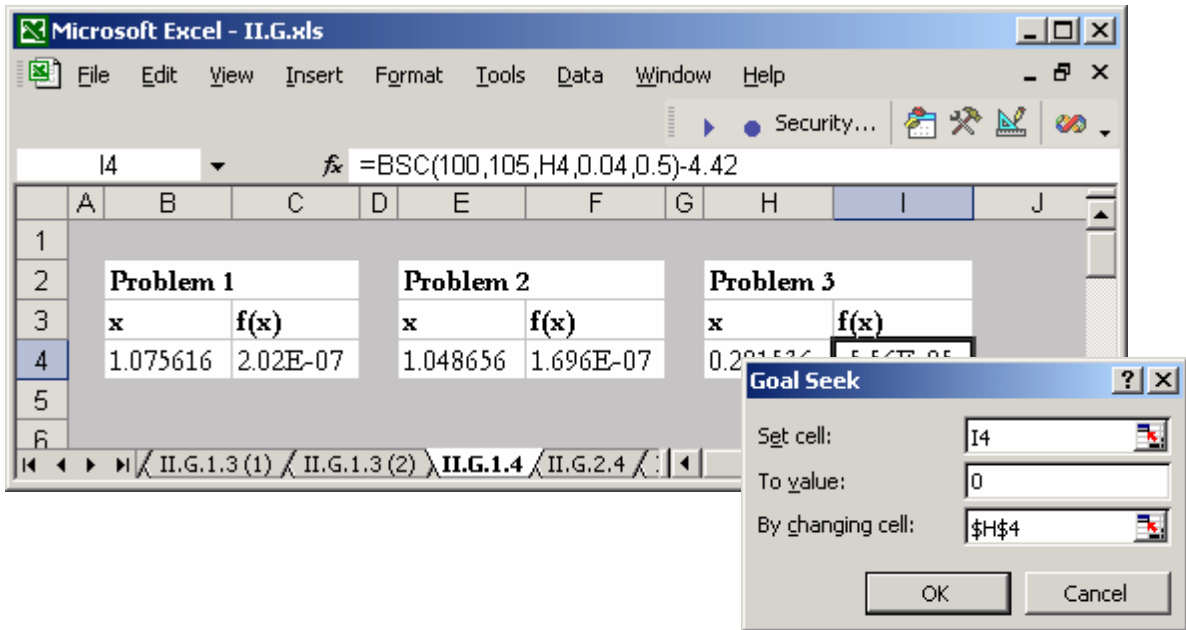
Figure II.G.6 shows that to two decimal places the correct answer is 4.87%. Note that starting with $x = 1$ does not work since $x = 1$ is itself a (false) solution (see Section II.G.1.1).

We will not pursue Newton–Raphson for problem 3 – the differentiation is too intricate.

II.G.1.4 Goal Seek

Excel's equation solver is called Goal Seek. It can be found under the Tools menu. It uses an iterative approach which is a modification of bisection. Below we apply it to all three problems. In the case of problem 3 the equation $f(x)$ is found in cell I4 (see Figure II.G.7). The equation solver will adjust the variable x (in cell H4) until I4 reaches the value zero (specified by the user in 'Goal Seek'). That is, the equation solver will change volatility to equate the Black–Scholes option price with the market price (equivalent to setting the Black–Scholes price less the market price equal to zero). The value of volatility implied by the market price is thus 20.1536% p.a. (under the assumptions of the Black–Scholes model). Note that in problem 3 we could simply have asked Goal Seek to set a cell containing '=BSC(100,105,H4,0.04,0.5)' to value 4.42 by changing cell H4.

Figure II.G.7: [Goal Seek](#)



II.G.2 Numerical Optimisation

We turn next to the topic of numerical optimisation. In Section II.G.1 we dealt with solving equations. An equation can always be manipulated into the form $f(x) = 0$, and the problem is to find which value(s) of x produce the answer 0 when operated on by the function. In optimisation we are concerned to find the value(s) of x which produce maximum or minimum function values, depending on the circumstances.

Analytical optimisation was covered in Section II.C.7. It involves finding the maximum or minimum of a function by finding points at which the function derivative is zero. In finance we encounter such optimisation problems in a number of contexts, for instance using maximum likelihood estimation for the parameters in a GARCH model (see Chapter II.F) and for fitting distributions to the returns to financial assets, calibrating option pricing models, solving optimal hedging problems and building optimal portfolios. We use *numerical optimisation* as an alternative to analytical optimisation either when the explicitly defined function to be optimised does not lend itself to those techniques, or when the function is not explicitly defined. Numerical optimisation is a vast subject area. We shall restrict ourselves to *gradient methods*, which get their name from the fact that they employ the (partial) derivatives of the function to be optimised.

II.G.2.1 The Problem

Numerical methods are often required in finance to optimise the value of something when that something depends on multiple inputs. Imagine we have a portfolio containing multiple assets. We could determine the optimal holding of asset 1 to minimise portfolio risk, assuming all other holdings constant. The problem with this univariate approach is that changes in the holdings of asset 1 will usually mean that the holdings of asset 2 need to be adjusted. In the portfolio setting the sensitivity of each asset to changes in its weights depends on what else is in the portfolio – that is, everything is interconnected. The interconnections between the variables necessitates multivariate optimisation, or optimising for all the variables simultaneously.

In Section II.D.2 we solved a multivariate optimisation problem. The problem was to construct a minimum risk portfolio for three assets having expected returns of 8%, 10% and 7% respectively. The covariance matrix of returns was:

$$\mathbf{V} = \begin{pmatrix} 0.01887 & 0.01722 & 0.00721 \\ 0.01722 & 0.02763 & 0.00604 \\ 0.00721 & 0.00604 & 0.00966 \end{pmatrix}.$$

An expected return of 9% was required and the investment was to represent 100% of the available funds. We solved the problem, firstly by constructing the Lagrangian and solving the resulting system of simultaneous equations, and secondly by using Excel's Solver facility.

Our objective here is to give some background regarding the mechanism behind Solver. We will end the section by investigating this problem subject to the additional constraints that there should be no short selling. In fact, whereas one would often use Solver, or some other numerical optimiser, for convenience, these software tools were developed to cope with situations in which

there is no explicit function to be optimised. There the only information may be function evaluations, each obtainable only at what might be considerable cost.

II.G.2.2 Unconstrained Numerical Optimisation

The Newton–Raphson method generalises to give ‘Newton’s method’. This amounts to solving the vector equation $\mathbf{g}(\mathbf{x}) = \mathbf{0}$, where \mathbf{g} is the gradient vector of partial derivatives of the function which is to be optimised. Thus we focus on finding a ‘flat’ spot, a point where the partial derivatives are all zero.

Example II.G.6: The gradient

If $f(x, y) = x^2 + 6xy + 2y^3$ (as in Example II.C.10) then the partial derivative of f with respect to x is $\frac{\partial f}{\partial x} = 2x + 6y$. Similarly, the partial derivative of f with respect to y is $\frac{\partial f}{\partial y} = 6x + 6y^2$. Note the

notation (a type of delta rather than d) that is used to indicate a *partial* derivative, i.e. $\frac{\partial f}{\partial x}$ rather than $\frac{df}{dx}$.

Thus at $(1, 1)$, for instance, the gradient is given by $\frac{\partial f}{\partial x} = 2 + 6 = 8$ and $\frac{\partial f}{\partial y} = 6 + 6 = 12$.

This gradient is expressed as a vector, i.e. $\mathbf{g}(1,1) = \begin{bmatrix} 8 \\ 12 \end{bmatrix}$. It indicates that in the x direction the rate of increase of the function is 8, and that the rate of increase in the y direction is 12. The maximum rate of increase from the point $(1,1)$ is in the direction $\mathbf{g}(1,1)$. In general, *for any function $f(\mathbf{x})$, the maximum rate of increase of the function’s value from the value $f(\mathbf{x})$ is in the direction $\mathbf{g}(\mathbf{x})$, where $\mathbf{g}(\mathbf{x})$ is the function’s gradient vector.*

Using the Taylor expansion (see Section II.C.3.3) of each component of \mathbf{g} , we may write:

$$\mathbf{g}(\mathbf{x} + \mathbf{h}) \approx \mathbf{g}(\mathbf{x}) + \mathbf{H}(\mathbf{x})\mathbf{h} \quad (\text{II.G.5})$$

where \mathbf{H} is the Hessian matrix of second partial derivatives (see Section II.D.3, for example).

Example II.G.7: The Hessian

For Example II.G.6, $\mathbf{H}(x, y) = \begin{bmatrix} 2 & 6 \\ 6 & 12y \end{bmatrix}$, so equation (II.G.5) becomes

$$\begin{aligned} \mathbf{g}(x + h, y + k) &\approx \mathbf{g}(x, y) + \begin{bmatrix} 2 & 6 \\ 6 & 12y \end{bmatrix} \begin{bmatrix} h \\ k \end{bmatrix} \\ &= \begin{bmatrix} 2x + 6y \\ 6x + 6y^2 \end{bmatrix} + \begin{bmatrix} 2h + 6k \\ 6h + 12yk \end{bmatrix} = \begin{bmatrix} 2x + 6y + 2h + 6k \\ 6x + 6y^2 + 6h + 12yk \end{bmatrix} \end{aligned}$$

So at (1, 1) we have

$$\mathbf{g}(x + b, y + k) \approx \begin{bmatrix} 8 \\ 12 \end{bmatrix} + \begin{bmatrix} 2b + 6k \\ 6b + 12k \end{bmatrix}.$$

In the Newton–Raphson method we wish to move from our present point, \mathbf{x} , to a new point, $\mathbf{x} + \mathbf{h}$, where $\mathbf{g}(\mathbf{x} + \mathbf{h}) = \mathbf{0}$. So the required translation, \mathbf{h} , may be approximated by a solution to the system of linear equations

$$\mathbf{g}(\mathbf{x}) + \mathbf{H}(\mathbf{x})\mathbf{h} = \mathbf{0}. \tag{II.G.6}$$

Example II.G.8: The translation \mathbf{h}

Continuing the above example, equation (II.G.6) gives

$$\begin{bmatrix} 8 \\ 12 \end{bmatrix} + \begin{bmatrix} 2b + 6k \\ 6b + 12k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ or } \begin{cases} 2b + 6k = -8 \\ 6b + 12k = -12 \end{cases}$$

This system has solution $b = 2, k = -2$. Thus we would move from (1, 1) to (3, -1) in the first iteration of the method.

The Newton–Raphson method is often written as

$$\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} - \mathbf{H}^{-1}\mathbf{g}(\mathbf{x}^{(n)}), \tag{II.G.7}$$

although in practice the inverse Hessian may not be explicitly evaluated (see below). Here the ‘step number’ in the iteration is written as a bracketed index, since a subscript would indicate a component of the vector and an unbracketed index might be confused with a power.

In practical situations only function evaluations are possible, and derivatives and second derivatives have to be approximated by obtaining function evaluations near to the current point in the search, $\mathbf{x}^{(n)}$. This becomes extremely expensive in terms of function evaluations, which equate to time and/or money, and sophisticated techniques have been developed which do not need approximations for the second partial derivatives. Instead the iteration incorporates a matrix which converges automatically to the inverse of the Hessian. These are called *quasi-Newton* methods, and Solver incorporates a quasi-Newton routine.

A physical analogue might help. Imagine that you are standing on a hillside in the fog, and that your objective is to reach the top of the hill. This is a two-dimensional optimisation problem. Your present location, $\mathbf{x}^{(n)}$, is fixed by two numbers, for example latitude and longitude. The function to be optimised is the height. You estimate the gradient by taking a step eastward and noting the change of height, and then a step northward and noting the change of height. The algorithm will then determine the direction in which you should step off, and how far you should

go in that direction, before you should stop and repeat the calculations. The algorithms are sometimes called *hill-climbing* routines.

Eventually you will arrive at a hilltop. Whether it is the top of the highest hill will not be clear. That is an unsolvable problem in multivariate optimisation which impinges significantly in areas such as parameter estimation in GARCH analysis. Here the data sets are huge. In such applications Excel’s Solver does not have adequate power and a bespoke software package is needed.

II.G.2.3 Constrained Numerical Optimisation

In unconstrained optimisation all possible values of the variables are available. In constrained optimisation this is not the case. For instance, in portfolio optimisation the sum of the portfolio weights cannot exceed 1, and must equal 1 if all funds are to be invested (see Section II.D.2). Furthermore, the weights must be constrained to be non-negative, unless short selling is permitted. Another common application of constrained optimisation is the calibration of an option pricing model to an implied volatility surface. Here we need to ensure that the model does not give negative prices or volatilities. And, of course, the maximisation of returns subject to some limits on risk is a constrained optimisation problem that is fundamental to all types of efficient resource allocation.

Consider the following very simple unconstrained univariate optimisation problem:

$$\text{maximise } y = -x^2 + 6x - 4.$$

We can solve this either by differentiating to find the turning point, or by using the algebraic technique of completing the square to obtain $y = 5 - (x - 3)^2$. Either way the maximum is at (3, 5).

Now consider the following two constrained optimisation problems:

A
Maximise $y = -x^2 + 6x - 4$
subject to $x \leq 2$

B
Maximise $y = -x^2 + 6x - 4$
subject to $x \leq 5$

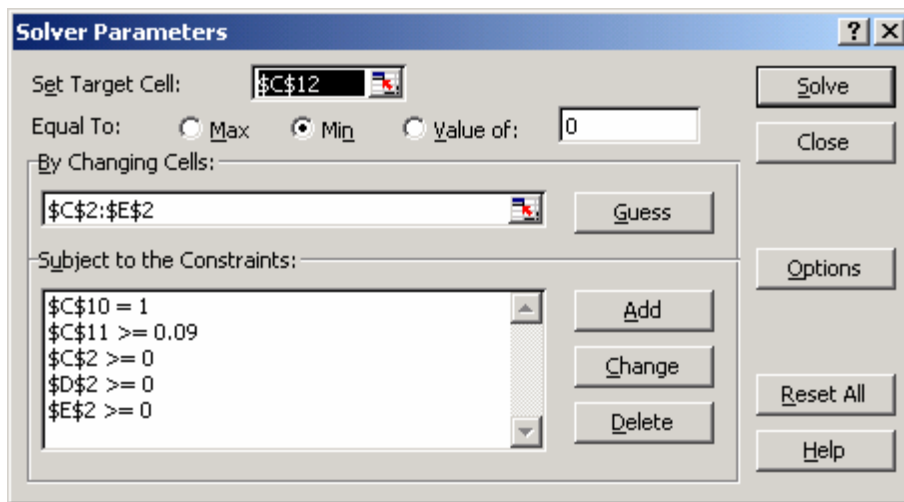
In the first case, problem A, the constraint is *active* and the solution is $y = 4$ at $x = 2$. In the second case, problem B, the constraint is *inactive* and the solution is the unconstrained optimum of $y = 5$ at $x = 3$. Thus one approach to constrained optimisation is to solve 2^n different equality constrained optimisation problems, where n is the number of constraints. This covers

all of the possibilities, each constraint being included or excluded in all possible combinations in the formulations. Each equality constrained problem can then be converted into an unconstrained problem by constructing the Lagrangian. The full analysis is known as a *Kuhn–Tucker* analysis.⁴ You will not need to implement such approaches yourself, but it is instructive to know that optimisation packages such as Solver use them.

II.G.2.4 Portfolio Optimisation Revisited

We close this section by applying Solver to the revised portfolio optimisation problem specified earlier in this section. We want an expected return of at least 9%, to invest 100% of available funds, and have no short selling.

Figure II.G.8: [An application of Excel solver](#)



⁴ The approach is OK for problems with a small number of constraints even though it is inefficient – for instance, in most cases many of the formulations will have feasible solutions. However, even with only three constraints eight problems are spawned, so it is not really a practicable option for a software package.

Another approach is embodied in what are known as active set methods. At each iteration a subset of the inequality constraints, together with all of the equality constraints, are active. The iteration then attempts to move from the current point, at which all of the active constraints are satisfied as equalities, to the best point in which all of the active constraints are satisfied as inequalities.

However, that best point might not be feasible since in moving towards it a boundary for one of the currently inactive constraints might be encountered. If that should happen then that constraint is added to the 'active' list, and the process is repeated.

If, on the other hand, the best point is feasible, then the values of the Lagrange multipliers are examined. If they are all non-negative then a solution has been reached. If any one is negative then the corresponding constraint may be excluded from the active set, and the process repeated.

	active	inactive	inactive
weights	0	0.666667	0.333333
covariance matrix	0.01887	0.01722	0.00721
	0.01722	0.02763	0.00604
	0.00721	0.00604	0.00966
expected returns	0.08	0.1	0.07
proportion invested	1	active	
expected return	0.09	active	
portfolio variance	0.016038		
volatility	0.12664		

The solution shown in Figure II.G.8 is to invest two-thirds of the funds in asset 2 and one-third in asset 1. The constraint that all funds must be invested is active, as is the constraint that the expected return must be at least 9%. The ‘achieved’ volatility is 12.7%.

II.G.3 Numerical Methods for Valuing Options

A differential equation is one involving values of both a function and a derivative or derivatives of that function. When there is more than one underlying variable then the derivatives are partial derivatives (see Section II.C.5) and the equations are *partial differential equations* (PDEs).

The Black–Scholes equation is a PDE that governs the dynamics of all derivative securities in a complete market where the underlying asset dynamics are a geometric Brownian motion. It is

$$\frac{\partial V}{\partial t} + rS \frac{\partial V}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} = rV, \tag{II.G.8}$$

where V is the option value and where the other variables are as in equation (II.G.3). When the underlying asset has constant volatility and we are valuing a European option, equation (II.G.8) has an analytic solution: the celebrated ‘Black–Scholes formula’ (also called the ‘Black–Scholes–Merton’ formula; see Section I.A.8.7). But when the underlying volatility is not constant, or for ‘path-dependent’ options,⁵ it is not possible to find a closed form solution to the Black–Scholes PDE. In this section we will show how to use binomial lattices and finite difference methods to value an American, and how to use simulation to value an Asian option.

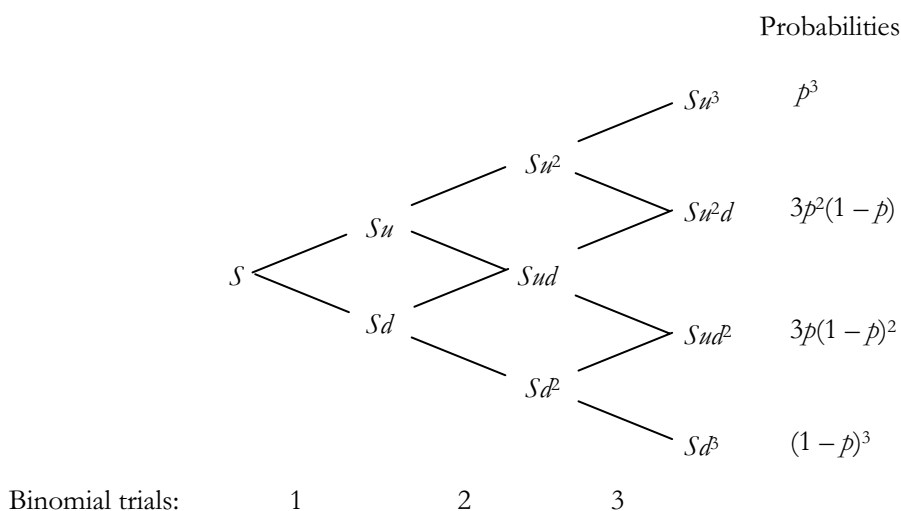
⁵ These are options whose payoffs depend on the path of the price of the underlying over a period of time. See Chapter I.B.9 for more details.

II.G.3.1 Binomial Lattices

Lattice methods (binomial or trinomial) are very popular for valuing options where there is some discretion regarding the timing of an event that affects its payoff. Binomial lattices and their applications to pricing and hedging options were introduced in Chapter I.A.8. Simple European options are normally valued using an analytic ‘Black–Scholes’ type formula. However, lattice methods may be used to value path-dependent options (such as American options, where the timing of exercise is at the holder’s discretion) or options based on an underlying with non-constant volatility. For a discussion of exotics that are valued using this method, see Chapter I.B.9.

In Section II.E.4.1 we showed how to model price movements of an asset using a binomial lattice. Part of that section is paraphrased below for your convenience. A security price, S , is assumed to move up by a factor of u ($u > 1$) or down by a factor of d ($0 < d < 1$). In the binomial lattice drawn in Figure II.G.9 this happens three times and therefore there are four possible outcomes. If r is used to count the number of upward movements (i.e. ‘successes’ in the Bernoulli trial – see Section II.E.4.1), then outcome Su^3 is the result of three successes, that is, $r = 3$. The outcome Su^2d is the result of two successes, so $r = 2$. The outcome Sud^2 is the result of one success so $r = 1$. Finally, Sd^3 is the result of zero successes, so $r = 0$.

Figure II.G.9 Binomial lattice



The binomial distribution gives the probabilities of each of these outcomes. The probability of achieving each outcome is given by ${}^nC_r p^r (1-p)^{n-r}$, where in the illustration $n = 3$, r is as defined above, and p has yet to be defined.

Cox, Ross and Rubinstein (1979) showed how such a model can be surprisingly effective, even for small n , in capturing a *lognormal* distribution for changes in asset prices.⁶ (See Section II.E.4.5 for a description of the lognormal distribution applied to underlying asset price changes.) We demonstrate this by using a six-period lattice to value a European put. The put has a life of 1 year. The current underlying asset value is 100 and the strike is 102. The volatility of the underlying is 20% and the risk-free rate is 5%. The Black–Scholes value is 6.45.

We first need to define u , d and p . We have to choose them in such a way that (approximately)

- the expected return on the asset over a time interval is $r\Delta t$, where r is the risk-free rate (we are using the principle of risk-neutral valuation here),
- the standard deviation of returns over a time interval is $\sigma\sqrt{\Delta t}$.

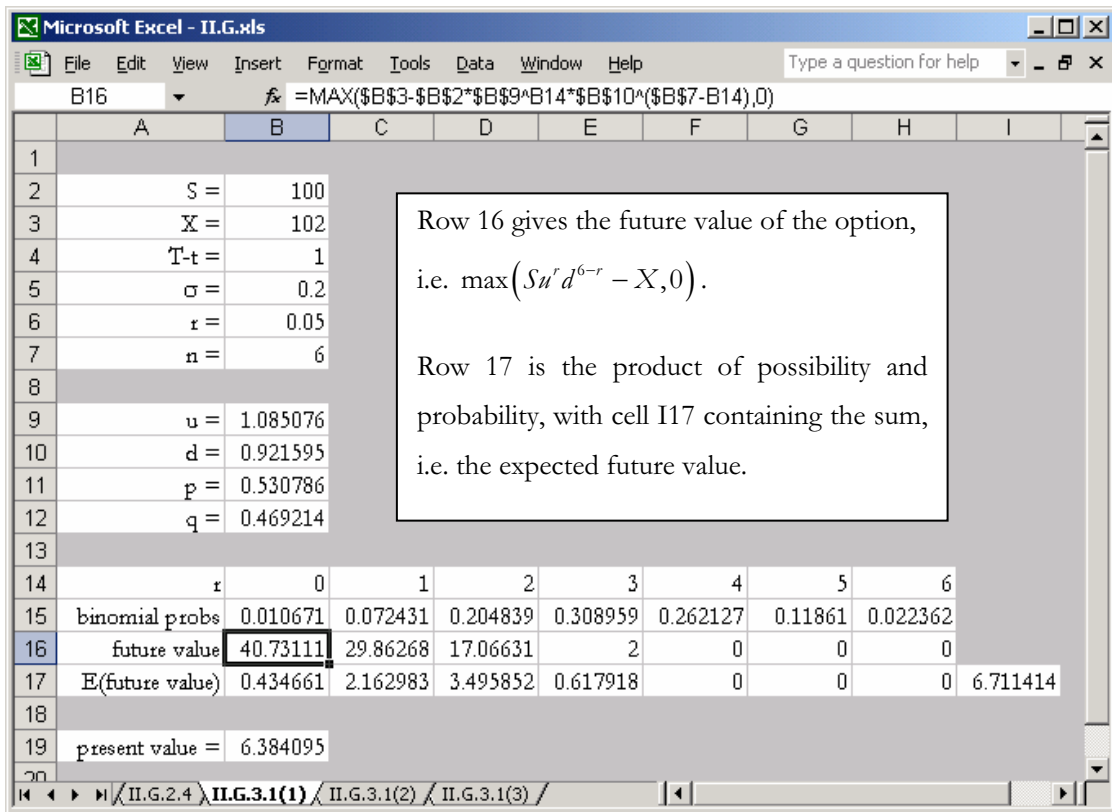
Thus there are three choices to make, u , d and p – and only two constraints. So there are many alternative parameterisations (or ‘discretisations’). We use below the Cox, Ross and Rubenstein parameterisation (see Figure II.G.10), which is used with continuous compounding. In this parameterisation d is taken to be $1/u$, with u and p given by

$$u = \exp\left(\sigma\sqrt{\frac{(T-t)}{n}}\right) = \exp\left(0.20\sqrt{\frac{1}{6}}\right) = \exp\left(\frac{1}{5\sqrt{6}}\right) \approx 1.085, \quad (\text{II.G.9})$$

$$p = \frac{\exp\left(r \times \frac{T-t}{n}\right) - d}{u - d} = \frac{\exp\left(0.05 \times \frac{1}{6}\right) - \exp\left(-\frac{1}{5\sqrt{6}}\right)}{\exp\left(\frac{1}{5\sqrt{6}}\right) - \exp\left(-\frac{1}{5\sqrt{6}}\right)} \approx 0.531. \quad (\text{II.G.10})$$

⁶ This is achieved through the combination of the multiplicative possibilities and the binomial probabilities, the latter tending to normal probabilities as the number of intervals increases.

Figure II.G.10 Binomial lattices using Excel (1)



Here the binomial value (6.384095) is quite close to the Black–Scholes value (6.45). We see that, even with only six steps, there is an error of just 1.1%, computed from $\left(\frac{6.45 - 6.384}{6.45} \times 100\right)$, in estimating the value of the option.

For an example of an alternative discretisation the reader may wish to compare that given in Section I.A.8.4, namely

$$u = 1 + \sigma\sqrt{dt}, \quad v = 1 - \sigma\sqrt{dt} \quad \text{and} \quad p = \frac{1}{2} + \frac{\mu\sqrt{dt}}{2\sigma}.$$

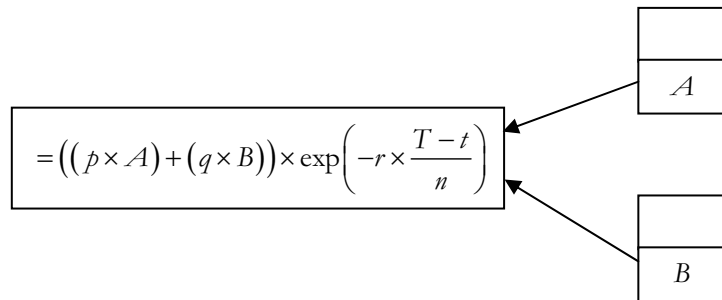
Using six steps this gives an estimated value of 6.58, which is in error by about 2%.

Of course, interesting though it is, there is no point in using numerical methods for valuing a European – the analytical Black–Scholes model does so far more efficiently. But what we can do is disaggregate the computation into the individual stages. We will then be able to adapt it to value an American.⁷ The layout is shown in Figure II.G.11. Each node in the lattice has an upper

⁷ An American is an example of a *path dependant* option. For a non path dependent option its value at any point in time depends on the value of the underlying and not on the past values of the underlying. For an American the path taken by past prices is relevant – early execution might have been triggered.

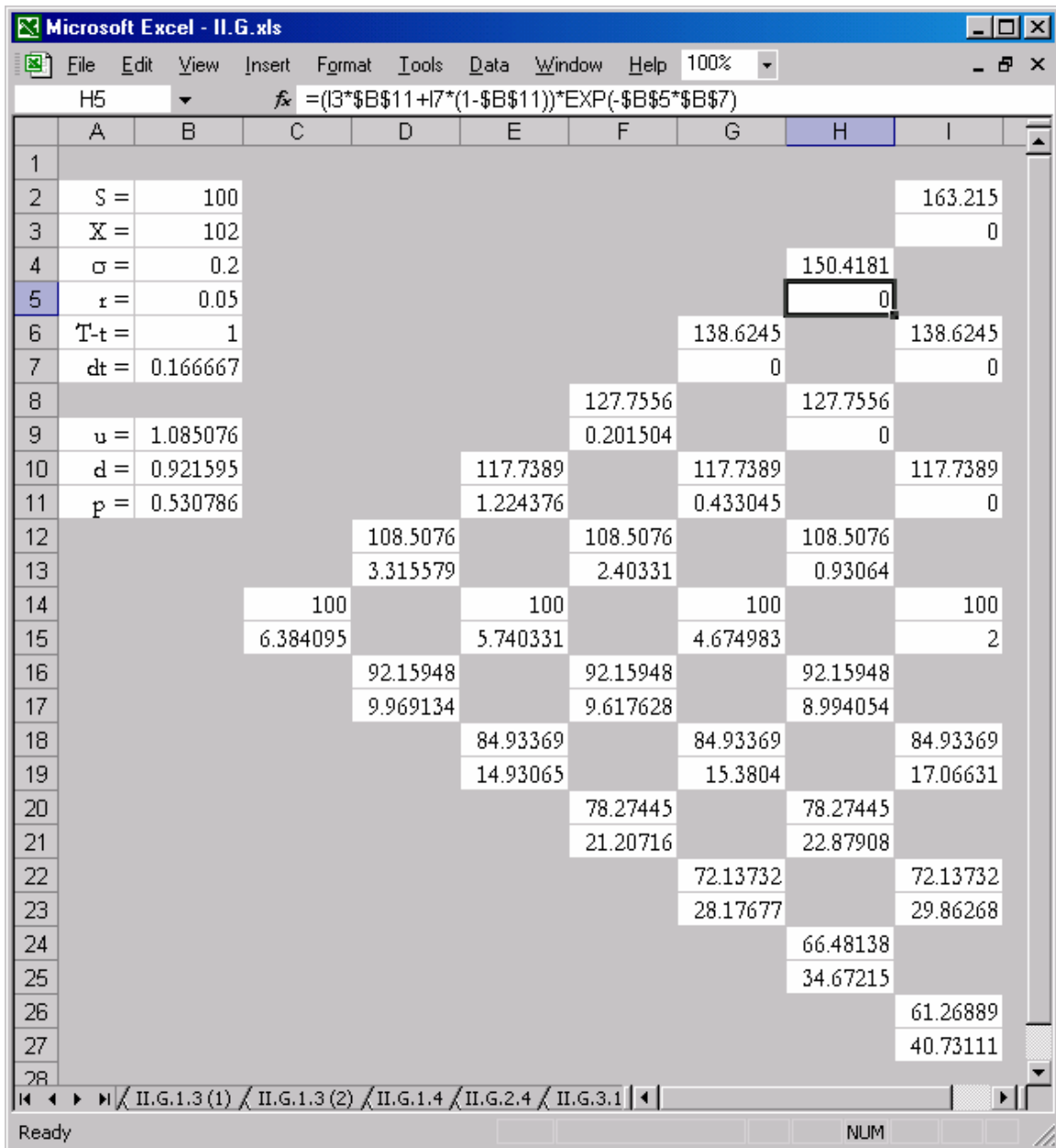
value computed by moving ‘forward’ through the lattice, as per Figure II.G.9. This is the stock price at that node, calculated using u and d . The stock price at D12, for example, is calculated as $Su = 100(1.085076)$. At F12, the stock price is calculated as $Su^2d = 100(1.085076)^2(0.921595)$, and so forth.

The lower value is computed in a backward pass. The lower values in the rightmost column are just the future values, $\max(Su^r d^{6-r} - X, 0)$. Having established the terminal option values, we then value the option at each of the penultimate nodes (see column H in Figure II.G.11) using the risk-neutral valuation method explained in Section I.A.8.4. To value the option at H5, we need to use the values previously calculated at I3 and I7, hence the need for a backward pass through the lattice. The process continues backward through the lattice until the current value of the option is determined. To continue backwards we compute a discounted expectation thus:

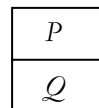


Notice that the value calculated in Figure II.G.11 (6.384) exactly matches the value calculated previously in Figure II.G.10.

Figure II.G.11: Binomial lattices using Excel (2)

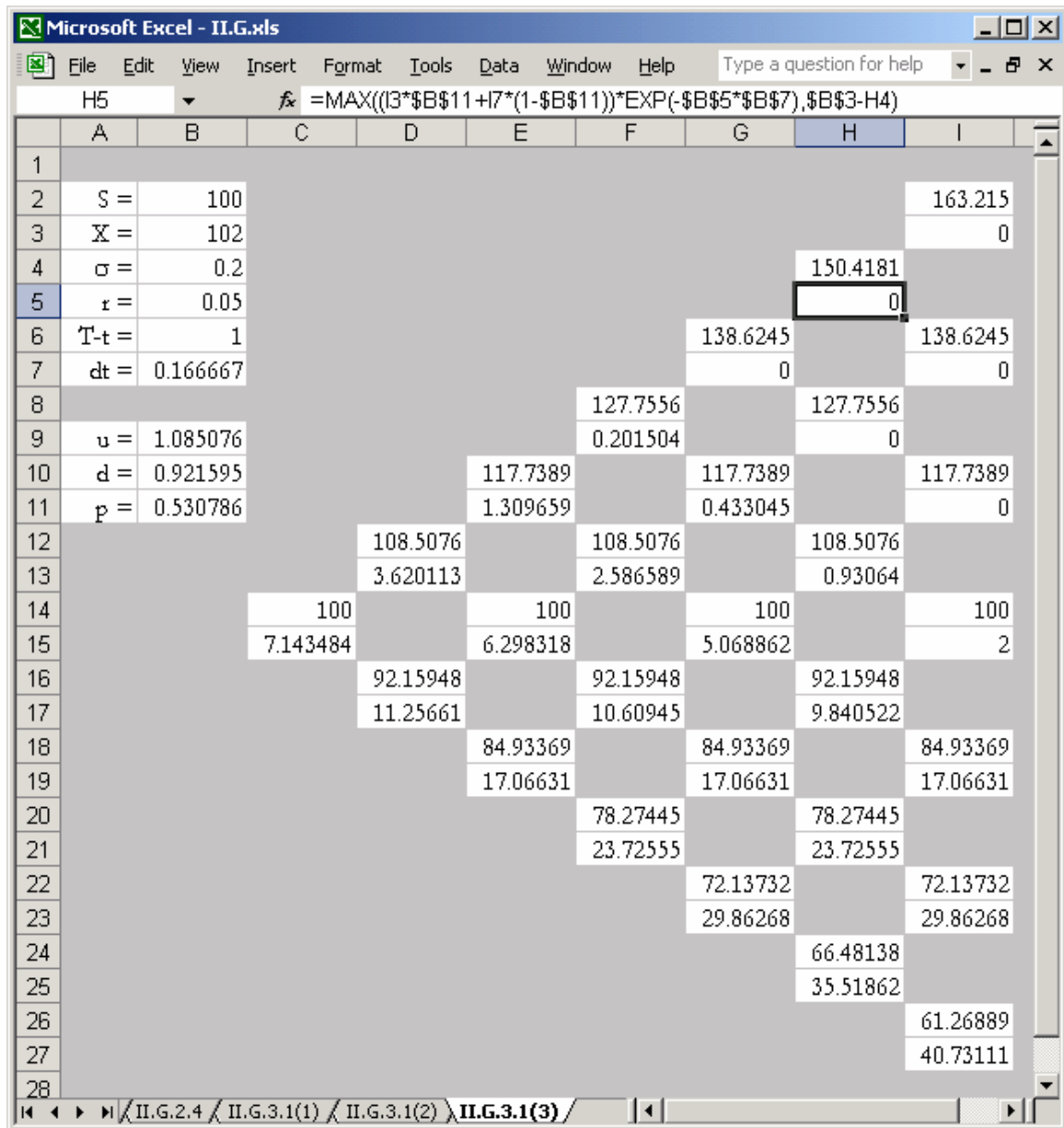


Finally, we are in a position to value the American call. All we need to do is modify the cells in the backward pass to allow for the possibility of early exercise. At every node, we consider whether to exercise, in which case the value of the option is just its exercise value, being $P - X$ for a call option (or $X - P$ for a put). Labelling a pair of cells thus:



we now replace Q by $\max(Q, (P - X))$ in every lower cell except for those in the last column, as shown in Figure II.G.12.

Figure II.G.12 Binomial lattices using Excel (3)



So we see that the estimated value of the American is 7.14, compared to a Black–Scholes value of 6.45 for the European. The lattice approach is ideally suited to American option valuation because the exercise decision will depend on future values. As the lattice method starts at the terminal option values moving to the present, it is easy to incorporate the exercise decision into the valuation process.

II.G.3.2 Finite Difference Methods

Most option values are determined by a PDE, such as the Black–Scholes PDE. These PDEs can become quite complex, particularly when there are many underlying risk factors. For instance, PDEs can become complex for the valuation of convertible bonds, where both equity and interest rate uncertainty play an important role, or for the valuation of currency protected equity options, where the foreign exchange rate is another risk factor that affects the price. For the solution of such a PDE, it may be that a lattice method gives unreliable results. Finite difference methods may provide more stable solutions to the PDE than any lattice method.

As in the previous section, we shall use the Black–Scholes PDE as an example. Recall from equation (II.G.8) that this is

$$\frac{\partial V}{\partial t} + rS \frac{\partial V}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} = rV,$$

and when σ is not a constant there is no analytic formula for V that satisfies this equation. In this section we show how to solve the equation for V by replacing the partial derivatives by approximations, called ‘finite difference’ approximations. Finite differences are so called because they are derived from the differences between function values at points (i.e. variable values) which differ by small but finite (not infinitesimal) amounts.

The fundamental relationship (written here in terms of an ordinary, non-partial derivative) is based on the fundamental relationship that is used to define a derivative in calculus, that is, that $\frac{dy}{dx} \approx \frac{\delta y}{\delta x}$. For any function $f(x)$ and for a small value of b , the first derivative $f'(x)$ is approximated by:

$$f'(x) \approx \frac{f(x+b) - f(x)}{b}. \tag{II.G.11}$$

This is called a (first-order) *finite difference approximation*.

A finite difference approximation to the second derivative can be obtained using f' in place of f :

$$\begin{aligned} f''(x) &\approx \frac{f'(x+b) - f'(x)}{b} \approx \frac{\left(\frac{f(x+2b) - f(x+b)}{b} \right) - \left(\frac{f(x+b) - f(x)}{b} \right)}{b} \\ &= \frac{f(x+2b) - 2f(x+b) + f(x)}{b^2}. \end{aligned} \tag{II.G.12}$$

We shall be using variants of these for our three partial derivatives. They are:

$$\frac{\partial V}{\partial t} \approx \frac{V(t + \delta t, S) - V(t, S)}{\delta t},$$

$$\frac{\partial V}{\partial S} \approx \frac{V(t + \delta t, S + \delta S) - V(t + \delta t, S - \delta S)}{2\delta S},$$

$$\frac{\partial^2 V}{\partial S^2} \approx \frac{V(t + \delta t, S + \delta S) - 2V(t + \delta t, S) + V(t + \delta t, S - \delta S)}{\delta S^2}.$$

Whilst these approximations are all valid, they do seem to be rather arbitrary. Why, for instance, is $t + \delta t$ used in several places where t would do? The reason becomes clear when we substitute these expressions into the Black–Scholes PDE (equation (II.G.8)). By making this substitution we produce a relationship linking $V(t, S)$, $V(t + \delta t, S - \delta S)$, $V(t + \delta t, S)$ and $V(t + \delta t, S + \delta S)$.

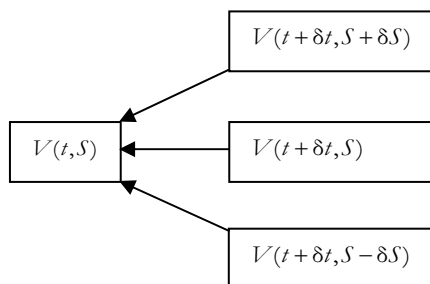
We can reorganise this to get:

$$V(t, S) = \frac{1}{1 + r\delta t} [aV(t + \delta t, S + \delta S) + bV(t + \delta t, S) + cV(t + \delta t, S - \delta S)]$$

where $a = \frac{S\delta t}{2\delta S} \left(\frac{S\sigma^2}{\delta S} + r \right)$, $b = \left(1 - \left(\frac{S\sigma}{\delta S} \right)^2 \delta t \right)$ and $c = \frac{S\delta t}{2\delta S} \left(\frac{S\sigma^2}{\delta S} - r \right)$. (II.G.13)

You need not worry about the details of this rather formidable expression. The point is that it gives an expression for $V(t, S)$ in terms of $V(t + \delta t, S - \delta S)$, $V(t + \delta t, S)$ and $V(t + \delta t, S + \delta S)$. This means that we can iterate backwards from known future values.

Diagrammatically:



This enables us to construct a spreadsheet which delivers today’s value by repeatedly using the relationship to compute backwards from the discounted, risk-neutral expiry value. Furthermore, we can add ‘IF’ statements to each cell to implement the valuation of path-dependent options in the same way as with binomial lattices.

We demonstrate with a spreadsheet to value the option from Section II.G.3.1, that is, a European put with a life of 1 year. The current asset value is 100 and the strike is 102. The volatility of the underlying is 20% and the risk-free rate is 5%. The Black–Scholes value is 6.45.

Figure II.G.13 [Finite differences using Excel](#)

	A	B	C							
24										
25				125.00	-8.02	5.87	-1.72	2.39	-0.08	1.26
26				120.00	5.27	-0.51	2.62	0.66	1.70	0.90
27				115.00	1.05	3.01	1.71	2.35	1.77	1.97
28	dS =	5.00		110.00	3.77	3.01	3.32	2.95	3.01	2.78
29	dt =	0.05		105.00	4.60	4.69	4.48	4.45	4.29	4.00
30	X =	102.00		100.00	6.51	6.39	6.34	6.23	6.15	6.04
31	σ =	0.20		95.00	8.72	8.68	8.62	8.57	8.51	8.45
32	τ =	0.05		90.00	11.49	11.48	11.47	11.46	11.45	11.44
33				85.00	14.78	14.83	14.87	14.92	14.97	15.02
34				80.00	18.58	18.67	18.78	18.88	19.00	19.11

Figure II.G.13 shows a small portion of the spreadsheet. It has been set up with 20 columns, so $\delta t = 0.05$ ($1/20$). δS has been chosen to be 5. The highlighted cell indicates that at time 0, if the underlying has value 100, then the option has an approximate value of 6.51.

In principle every cell of this spreadsheet has meaning. For instance, the entry to the right of and below the highlighted cell indicates that at time 0.05, if the underlying has value 95, then the option has an approximate value of 8.68. However, there are difficulties here concerned with convergence and numerical instabilities. It seems obvious, for example, that a finer grid would be helpful, particularly with respect to the values of the underlying. The spreadsheet [II.G.13](#) has been set up to facilitate this by changing the value of δS . However, reducing S leads to instability owing to the build-up of numerical errors. This effect can already be seen with the present value of δS for values of S of 120 and above. To counter this it is necessary to decrease δt . In fact it is necessary to decrease δt by the square of the factor by which δS is decreased. This is awkward with a spreadsheet, since it means rebuilding the sheet to accommodate more columns within the same time period, so bespoke software is better.

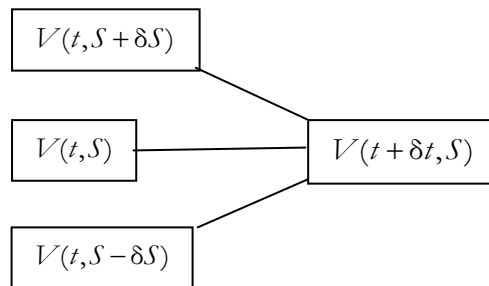
Better still is to use more sensible finite difference approximations, such as

$$\frac{\partial V}{\partial t} \approx \frac{V(t + \delta t, S) - V(t, S)}{\delta t},$$

$$\frac{\partial V}{\partial S} \approx \frac{V(t, S + \delta S) - V(t, S - \delta S)}{2\delta S},$$

$$\frac{\partial^2 V}{\partial S^2} \approx \frac{V(t, S + \delta S) - 2V(t, S) + V(t, S - \delta S)}{\delta S^2}.$$

The difficulty here is that these lead to a linkage between V values, which looks like this:



This means that we cannot explicitly write an earlier value in terms of later ones. However, if we consider all such relationships, together with the fact that we know what the option values will be for two extreme values of S , then we will have a large system of linear equations which we can solve to find earlier values of V from known later values of V . There are sophisticated software tools to do this – it is not to be attempted by the amateur! The resultant bespoke software is powerful and flexible, and represents a very useful analytical tool.

II.G.3.3 Simulation

Chapter I.B.9 on exotic options discusses the situations in which simulation is recommended for option valuation (see especially Section I.B.9.14). Monte Carlo simulation is most efficient in situations where the payoff is path-dependent, for example where it depends on the maximum or average price of the underlying. Monte Carlo simulation can be used in general for any discrete path-dependent payoff, such as average or reset options.

We showed in Section II.D.4.2 how to simulate correlated returns.⁸ In this section we will simulate the random walk of the price over time of the underlying asset from Section II.G.3.1 above. We will then use that simulation to value a path-dependent option – an Asian (see Section I.B.9.10). This is a call with a life of 1 year, the strike being defined as the arithmetic mean of the underlying price at defined points of time after 3 months, 6 months, 9 months and at expiry.

In Section II.D.4.2 we simulated continuously compounded returns, which we assumed to be normally distributed. We had information about those returns from which to estimate the mean and standard deviation. When we are simulating prices given the drift rate and volatility we need the formulae which give the mean and standard deviation of the returns in terms of the drift rate and volatility. Those formulae depend on the time interval being simulated. They are:

$$\text{return mean} = \left(\mu - \frac{\sigma^2}{2} \right) t \tag{II.G.14}$$

$$\text{return standard deviation} = \sigma\sqrt{t} \quad (\text{II.G.15})$$

where μ is the drift rate, σ is the volatility and t is the time period to be simulated. However, to value the option we need to perform the simulation in a risk-neutral environment. Unless we do this we will not know what discount rate to use to compute the present expected value from the future expected value. To achieve this we use r , the risk-free rate of return, instead of μ , the drift rate.

We give in Figure II.G.14 the output from a *pilot simulation*, based on 20 repetitions. We see from this that the expected value of the option is 5.52. However, pressing the F9 button gives us a different set of 20 realisations, and a different mean. Doing so repeatedly produced the following string of means: 5.52, 4.69, 2.63, 6.08, 2.32, 5.82, 3.77, 3.27, 6.76 and 5.89. These vary quite significantly, and this is only to be expected. Individual outcomes vary even more significantly, with present worth varying in Figure II.G.14 from $0.00 \times e^{-0.05} = 0.00$ to $22.48 \times e^{-0.05} = 21.38$. What we are trying to find is the expected value, and to do this we need many repetitions. Clearly 20 repetitions are not sufficient for us to be confident in our estimate. So the question is, how many repetitions do we need? To answer this we need to draw on the confidence interval and hypothesis testing work in Chapter II.F.

⁸ We used the Cholesky decomposition of the returns covariance matrix to simulate correlated returns. We have no need for such considerations here since we are concerned with just a single asset.

Figure II.G.14 Simulation using Excel

	A	B	C	D	E	F	G	H	I	J
1				S(0.25)	S(0.5)	S(0.75)	S(1)	fut opt val		mean
2	S(0) =	100		90.23	80.54	80.28	73.12	0.00		5.52
3	r =	0.05		93.23	91.54	110.33	127.03	17.60		
4	σ =	0.2		105.73	136.25	125.95	110.38	0.00		s.d.
5	t =	0.25		95.66	95.36	91.92	106.10	8.84		7.2985
6				104.03	125.26	113.97	137.62	17.27		
7	(r-σ ² /2)t =	0.0075		112.91	111.12	117.68	118.89	3.39		
8	σ sqrt(t) =	0.1		115.12	119.12	106.63	114.26	0.48		
9				97.08	108.67	114.19	114.98	6.25		
10				96.64	102.78	94.58	91.43	0.00		
11				94.66	100.32	108.76	101.05	0.00		
12				104.26	90.62	101.08	94.09	0.00		
13				104.43	117.15	142.59	151.37	22.48		
14				100.02	103.22	99.58	101.25	0.23		
15				90.04	90.26	91.76	106.46	11.83		
16				97.06	91.72	95.05	109.71	11.33		
17				114.93	124.73	107.67	94.58	0.00		
18				96.16	98.79	89.33	91.24	0.00		
19				106.97	100.94	107.16	96.82	0.00		
20				102.75	97.44	101.68	107.42	5.09		
21										

We first need to decide on the combination of accuracy and confidence that we require. Let us suppose in this instance that we want to estimate the true value to within an accuracy of 0.01 with 95% confidence. That amounts to saying that we wish to construct a 95% confidence interval of width 0.005, and to do that we need to know the standard error of the mean (see Sections II.F.4 and II.F.5). This is given by:

$$\text{standard error of mean of } n \text{ items} = \frac{\sigma}{\sqrt{n}} \tag{II.G.16}$$

Of course, we do not know σ , but we do have an estimate of it from our pilot survey, which shows $s = 7.2985$. Furthermore, repeatedly pressing F9 also gives repeated values of s , that is, repeated estimates for σ , and we can choose a reasonable upper bound by so doing. In this instance we shall choose 10, by which we mean that the true value of σ is of that order of magnitude, and probably somewhat less.

The final step in constructing a confidence interval is to argue that, for large values of n , the sample mean is approximately normally distributed. This is a consequence of the central limit theorem (see Section II.E.4.4), and follows even though the original distribution is a mixture of a

continuous set of possibilities together with the single possibility zero, which has a relatively high probability. So our 95% confidence interval will have a half width of $1.96 \times \frac{10}{\sqrt{n}}$, and we require

this to be 0.05. This gives $n = \left(\frac{19.6}{0.05}\right)^2 \approx 154,000$. This would be rather ambitious for Excel, and

highlights the fundamental problem with simulation – extensive computing power is required to achieve in quick time the large number of repetitions needed for accurate estimates. We will content ourselves with 5000 repetitions. According to our analysis this will give us a confidence interval of half width about

$$1.96 \times \frac{10}{\sqrt{5000}} \approx 0.3.$$

Implementing this and repeating 10 times gave the following string of estimates: 4.87, 5.05, 4.91, 5.03, 4.99, 5.00, 5.13, 4.92, 4.83 and 4.97.

The problem of needing many repetitions has been ameliorated by increasing computer power. It is also tackled by a battery of techniques known as *variance reduction techniques*. In one of the simplest, *antithetic variables*, whenever a simulation is run using a particular random variable, r , it is run again using $1 - r$. The results are then averaged. Thus there are twice as many repetitions, which would ordinarily reduce the variance of the estimator by a factor of $\sqrt{2}$. But the negative correlation between pairs of results introduced by using r and $1 - r$ produces a greater reduction in variance.

II.G.4 Summary

Much of mathematical finance hinges on the availability of computing power to solve otherwise intractable problems. It is in the nature of such problems that the greater the complexity, the more difficult it is to deliver cut-and-dried software solutions. The user therefore needs to have some knowledge of how the tools work, the better to be able to use them.

References

Cox, J C, Ross, S A, and Rubinstein, M (1979) 'Option pricing: a simplified approach', *Journal of Financial Economics*, 7, pp. 229–263.

III.A.1 Market Risk Management

Jacques Pézier¹

III.A.1.1 Introduction

What is market risk and whom does it concern? What do we mean by market risk management, and what does a market risk manager do in a day at the office? These are not theoretical questions with only right or wrong answers, they are practical questions that every financial as well as non-financial firm must grapple with; what appear to be reasonable answers depends very much on the activity, environment, culture, objectives and organisation of each firm.

In this chapter students are introduced to the four major tasks of risk management applied to market risks, namely, the identification, assessment, monitoring and control/mitigation of market risks. The difficulties faced in carrying out these tasks vary according to businesses. We therefore examine three typical activities – fund management, banking and manufacturing – to illustrate a broad spectrum of problems and state-of-the-art approaches. We aim to develop a conceptual and largely qualitative understanding of the topic. More detailed quantitative analyses are given in subsequent chapters.

III.A.1.2 Market Risk

To facilitate the analysis and understanding of risks faced by financial firms it is common practice to classify them into major types according to their main causes. Thus, banking risks are typically classified as being either market, credit or operational in origin. Broadly speaking, market risk refers to changes in the value of financial instruments or contracts held by a firm due to unpredictable fluctuations in prices of traded assets and commodities as well as fluctuations in interest and exchange rates and other market indices.² It is not clear to what extent market risks should be or can be considered for less liquid assets such as real estate or banking loans. Accountants usually shy away from attributing ‘fair values’ to such assets in the absence of reliable and objective market values, and consequently market risks are difficult to assess on illiquid assets. However, when the core activity of a business is to hold portfolios of illiquid assets, it would be dangerous to ignore their potential change in value.

¹ Visiting Professor, ISMA Centre, University of Reading, UK.

² By contrast, credit risk refers to changes in value of assets due to changes in the creditworthiness of an obligor and, at the limit, losses due an obligor failing to meet its commitments; operational risk is defined as the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events.

Banking supervisors have taken a special interest in codifying risks and in setting standards for their assessment. Their purpose is essentially prudential: to strengthen the soundness and stability of the international banking system whilst preserving fair competition among banks. To this end they have designed a set of minimum regulatory capital requirements for all types of traded assets, including some illiquid ones, based on detailed definitions and assessments of risk.³ By and large, banking supervisors have erred on the side of objectivity rather than comprehensiveness when defining market risks. Banks must allocate their assets to either a banking book or a trading book; market risks, bar exceptional circumstances, are recognised only in the trading book. A trading book consists of positions in financial instruments and commodities held either with a trading intent or to hedge other elements of the trading book; all other assets (e.g. loans) must, by default, be placed in the banking book. To be able to receive trading book capital treatment for eligible positions, some further basic requirements must be met such as clearly documented trading policies, daily mark to market (or mark to model) of positions and daily monitoring of position limits. To remain prudent, banking regulators have always inflated the capital treatment of credit risks in the banking book to cover for hidden market risks.

III.A.1.2.1 Why Is Market Risk Management Important?

Banking supervisors also hope that regulations will promote the adoption of stronger risk management practices, which they view as a worthwhile goal.⁴ Working in collaboration with the industry, they have certainly put these issues in the limelight and progress has been made. However, the development and adoption of better risk management practices in banks remains an objective ultimately beyond the reach of banking supervisors. It requires enlarging the purpose of risk management from a purely prudential objective (setting a limit on insolvency risks) to a broader economic objective (balancing risks and returns); it also requires enlarging the scope of market risk assessment to those areas that have been largely ignored by regulators because accrual accounting practices hide the risks and/or the risks are difficult to quantify objectively.

Outside financial services there are no prudential regulations offering guidelines for the management of market risk, but market risk nonetheless remains a major determinant in the success or failure of most economic activities and the welfare of people in free market economies. Suffice to observe how variations in the price of energy affect manufacturing and transportation, changes in interest rates affect the cost of mortgages and thereby property prices,

³ See Basel Committee on Banking Supervision (BCBS, 1996, 2004a). Insurance companies are subject to different solvency tests; harmonisation of insurance company solvency tests with minimum capital requirements for banks is a long-term aim for regulators. Pension funds and other funds designed to meet strict liabilities are also subject to solvency tests by the relevant regulatory authorities.

⁴ See BCBS (2004a, paragraphs 4 and 720).

and the performance of securities markets affects pensions. The absence of prudential regulations for non-financial firms gives an opportunity to reconsider the best way to recognise and tackle market risks. And what should become clear is that, satisfying as it may be to categorise risks according to causes, there is no classification system by which every risk would fall into one causal category and one category only, nor any management system that could control one type of risk without affecting others.

III.A.1.2.2 Distinguishing Market Risk from Other Risks

Some examples will illustrate how market, credit and operational risks are interrelated. On a macro-economic scale, consider the technology bubble that burst at the turn of the millennium. In 1996 Alan Greenspan, the Fed Chairman, described as 'irrational exuberance' the expectations placed on new technologies; indeed, they did not perform as well or as rapidly as predicted – a business or operational risk problem. A wave of corporate failures followed – a credit risk – and the Fed as well as many other monetary authorities across the world reacted by lowering interest rates – a market risk for bond portfolio holders. For an example on a micro-economic scale, consider a firm exposed to exchange-rate fluctuations – a market risk. It may seek cover by entering into a forward exchange-rate agreement with a bank, but it thereby takes a credit exposure on the bank if the bank has to pay the firm under the contract. Or consider a bank that makes a floating rate loan to a firm, thus taking primarily a credit risk on the firm; the bank may seek some degree of credit cover by asking for securities or property to be placed as collateral, but the value of the collateral will be subject to market risk.

Distinguishing market risks from other risks and managing them separately from and independently of other risks and profit considerations is therefore only valid up to a point. In any organisation, a balance must be struck between the degree of specialisation of risk management functions, so that market risks can be distinguished from other risks and managed separately, and the degree of interaction and coordination between functions so that they can operate coherently.

III.A.1.3 Market Risk Management Tasks

The Basel Committee on Banking Supervision has given a great deal of thought to the role and organisation of a risk management function. It distinguishes four main tasks that it defines as identification, assessment, monitoring and control/mitigation of risks.⁵ These tasks are relevant for market risks, as they are for most other risk types and for non-financial institutions as well as for banks.

⁵ See publications on the New Basel Accord, in particular BCBS (2004a, paragraphs 725–745 and 663(a)).

Identification is the necessary first step, but it may be less obvious than first thought. Real-world problems do not come neatly defined as in textbooks; they have to be recognised and delineated. Exposures to market risks can easily be overlooked because of either over-familiarity (risks we have always lived with without doing anything about them) or, at the other extreme, ignorance of new risks. The first case is all too common; for example, many corporates do not hedge currency exposures because they are not sure how to assess them or how to hedge them. In the end, it may be easier for a corporate treasurer to remain passive and blame the currency markets for a loss than to be active and have to explain why a loss was made on a hedge, the two circumstances having about equal probabilities. On the other hand, lack of familiarity may lead to over-cautious reactions. For example, investing in foreign equity markets, even when they are denominated in the same currency as the domestic market, is generally considered more risky than investing in the domestic equity market. But not recognising a new risk or combination of risks may be the greatest danger.⁶ The managers of the hedge fund LTCM, including two Nobel prize laureates in economics and finance, knew almost everything that could be known about market risks, but they were caught unawares by an unusual combination of events: the repercussions of the Russian bond crises of August 1998, diminished liquidity in major bond markets because of the withdrawal of a large market maker, and difficulty in raising new funds having just returned some capital to shareholders.

Assessment is the second step. The word initially chosen by the BCBS was ‘measurement’, but it has wisely been replaced by ‘assessment’ in more recent publications. Indeed, risks are not like objects that can be measured objectively and accurately with a simple measuring tape. Risks are about future *unexpected* gains or losses. The term ‘assessment’ reflects the need for a statistical model, that is, a coherent and relevant set of assumptions and parameters, some being supported by past evidence (e.g. former loss events) and others being chosen for the purpose of the exercise (time horizon, trading strategy, etc.). Banking supervisors have set qualitative and quantitative standards for the assessment of market risks to suit their aim, namely, the determination of prudent minimum capital requirements for banks. But internally, banks and other firms should take a wider view to choose standards that suit their own situation and objectives. In the next sections we shall illustrate how different assessment standards may be suitable for different businesses lines. But we shall not go into detailed quantitative techniques; these are covered in the following chapters.

⁶ Readers may remember how the US Secretary of Defence, Donald Rumsfeld, was once derided for trying to explain at a press conference that there are knowns and unknowns and that among the unknowns there known unknowns and unknown unknowns, the latter being the most dangerous. This is in fact old military lore: in the US Navy unknown unknowns are colloquially called ‘unk-unks’ and said to rhyme with sunk-sunk. Of course, Socrates had put this point across more elegantly when he said: ‘Wisest is the man who knows what he does not know.’

Monitoring refers to the updating and reporting of relevant information. Exposures and results can be monitored. Risks themselves cannot be monitored, but they should be frequently reassessed. Firms, strategies, markets, competition, technology and regulation all evolve, and therefore market risks also change over time. Even in a steady-state situation more information can be collected over time to develop a better understanding of market risks. Monitoring is particularly important when hedging strategies are in place so that one can verify the efficiency of these strategies and update the corresponding risk models.

Finally, *control* has been replaced by *control/mitigation* in the latest publications of the BCBS. Control gives too much the impression that market risks are intrinsically bad and therefore must be subject to limits, and a key responsibility of the risk manager is to verify that limits are not exceeded or, if they are, to blow the whistle. Mitigation has a wider meaning than control; it indicates that (i) there is a trade-off between risk and return and an optimal balance should be sought, and (ii) there are ways to manage market risks actively that need to be investigated.

How each of these tasks should be carried out by market risk managers and what specific problems they may encounter depend upon the business at hand. It would be ineffectual to give general answers to these questions. Rather, we shall explore three business types and show how the three tasks of market risk identification, assessment and control/mitigation vary between them.⁷ But first a few comments about how the risk management function should be organised

III.A.1.4 The Organisation of Market Risk Management

Banking regulators have put forward a few general recommendations for the organisation of the risk management function. They reflect a general consensus in the banking industry and are probably valid as well for many other businesses.⁸

- (i) The risk management function should be part of a risk management framework and policies agreed by the board of directors. The board and senior managers should be actively involved in its oversight.
- (ii) The risk management function should operate independently of the risk/profit-generating units; in particular, it should have its own independent sources of information and means of analysis. The risk management function should be given sufficient resources to carry out its tasks with integrity.
- (iii) The risk management function should produce regular reports of exposures and risks to line management, senior management and the board of directors. Non-

⁷ We leave aside in this chapter the more routine and easily understood 'monitoring' task.

⁸ The following four bullet points are not direct quotes but a summary by the author of recommendations that have appeared in various BCBS publications.

compliance with the risk management policies should be communicated immediately.

- (iv) The risk management process should be well documented and audited at regular time intervals by both internal and external auditors.

It is clear from these recommendations that the risk management function should be separate from and independent of risk-taking line management functions in the front office and support functions in the back office but should be in close communication with them. This is why most banks locate the risk management function in a separate ‘middle office’. The middle office must receive information on exposures from the front office in a timely fashion, but it should use its own independent information sources for prices and derived parameters such as volatilities and correlations, and it should use its own models to assess and forecast market risks. The middle office should produce regular (at least daily for banks) market risk reports for the front office and for senior management, containing detailed and aggregate risk estimates and comparisons against limits. These market risk reports are usually combined with credit exposure reports and profit attribution analyses where exceptional gains or losses as well as potential risks are explained. These reports must be immediately verified and approved by designated front office and senior managers. The middle office is also often responsible for producing statutory risk reports for banking supervisors.

The risk management function in financial firms is also normally in charge of preparing market risk management policies – to be submitted for the approval of the board – and of designing, assessing and recommending hedging strategies. It may also be required to calculate provisions and deferred earnings, to design and implement stress test scenarios, to establish controls and procedures for new products and to verify valuation methodologies and models used by traders. In a few instances, market risk managers may be asked to implement global hedging policies that would not sit naturally within any existing department. More recently, market risk managers have been asked to contribute to the analysis of the optimal level and structure of their firm’s capital (as compared to the evaluation of minimum regulatory capital) and to risk budgeting (also called ‘economic capital allocation’) with the aim of improving risk-adjusted return on capital.

Proper resources, independence and lines of communication to senior management and ultimately to an executive director on the main board are crucial for the integrity, credibility and efficiency of the risk management function in financial institutions that aim to derive a profit from taking market risks. Supporting and empowering the risk management function are lesser problems in firms that do not seek market risks but would rather avoid them. However, unless

there is a clearly defined and adequately supported market risk management function in these firms, market risks may not be properly appreciated and managed.

III.A.1.5 Market Risk Management in Fund Management

III.A.1.5.1 Market Risk in Fund Management

The core activity of fund management is to take market risks⁹ with the expectation of generating adequate returns.¹⁰ In most countries traditional funds are subject to strict regulations (about the type of securities in which they can invest, disclosure requirements, etc.) that are aimed at protecting investors. That is particularly so for pension funds that should provide long-term security to their members and, in return, enjoy certain tax advantages. Fund managers also often choose to limit and specialise themselves further according to market sectors or investment strategies, leaving to investors the choice to allocate their savings among funds and manage their own portfolio diversification; for example, specialist funds may describe themselves as ‘UK equity’, ‘capital guarantee’ or ‘index tracker’ funds.

But specialisation and constraints can only limit potential returns so, starting in the 1970s, private investment pools were created to offer to wealthy individuals, who might be more willing to take risks, the promise of greater returns by avoiding traditional fund constraints and regulations – for example, they can use short sales, derivative products and leverage. Interestingly, these new, unfettered funds became known as ‘hedge funds’ because many of them used trading strategies based on spreads between long and short positions rather than pure directional bets. Their popularity grew in the mid-1990s and even more so when the technology bubble burst and traditional funds’ returns inevitably tumbled. Total assets under hedge fund management may soon pass the trillion dollar mark.¹¹

Funds take market risks for the potential benefit of their investors. Fund managers themselves are only indirectly affected by market losses. Their income is usually a set percentage of the value of assets under management (plus some participation in profits in the case of hedge funds); it is not directly affected by losses. The market risks are born by the investors. But the reputation of fund managers and therefore their ability to retain existing investors and to attract new ones depends on their ability to manage their risks (and their clients). It is therefore crucial that (i) fund managers explain to their clients the risks they are taking, (ii) clients agree formally the terms

⁹ Funds also take credit and other risks but, bar special cases (e.g. a fund investing in a few high-yield corporate bonds or a certain emerging market), most funds invest in a large number of liquid, good-quality securities and therefore credit risks are less important than market risks.

¹⁰ In some cases returns must be sufficient to meet certain liabilities (e.g. pensions); in others, the fund managers’ objective will be to maximise some risk-adjusted performance measure.

¹¹ By comparison, in mid-2004, total worldwide bond and equity markets were valued at about \$70 trillion; more than half these assets were managed by institutions. In the USA, mutual funds alone managed about \$7 trillion of assets.

and conditions of their investment, and (iii) fund managers keep to the terms that have been agreed.

III.A.1.5.2 Identification

It should be an easy task for fund managers to identify market risks because they normally have chosen deliberately to take those risks. Nonetheless some risks may be overlooked, especially in funds following sophisticated strategies. In general, funds following spread or arbitrage type strategies will have reduced primary directional risks but will have increased exposures to secondary risks. For example, if a fund is allowed to short securities, the uncertainty in the repo cost incurred over the long term may be quite considerable, as well as difficult to estimate. Likewise, the spread between two similar securities, say A and B shares of a company, may be drastically affected by legal, tax or regulatory changes. It would not be much consolation to decide that such events are operational rather than market risks if they have not been foreseen.

Liquidity risk is another relevant concern, particularly for hedge funds.¹² Some assets may not be bought or sold at the anticipated price because the transaction is too large compared to the market appetite. Traditional funds are bound by regulations to hold only highly liquid positions. Hedge funds, on the other hand, do not have such constraints and may end up holding relatively large positions in specific securities. If in addition they are highly leveraged, they can easily be thrown into a momentary cash-flow squeeze or even a terminal problem by a liquidity crisis. We have already referred to LTCM as an example.

When funds have to meet specific liabilities – and many do¹³ – managers seek to maintain a stable surplus of assets over liabilities and should therefore be concerned by possible market risks on the liability as well as on the asset side. Actuarial practices and accounting standards have generally overlooked or hidden these risks in the past but new rules are now coming into effect that bring them to the fore. For example, in the United Kingdom, Financial Reporting Standard 17 (FRS 17) prescribes¹⁴ that asset and liabilities in company pension schemes be immediately

¹² Liquidity risks deserve to be analysed separately from market risks, but the two are closely related. The liquidity of a security can be characterised by its average daily trading volume. Usually, the bid–offer spread increases rapidly with the size of a transaction relative to the average daily trading volume when that fraction is significant. Exceptionally, there are securities that do not trade regularly and yet can be traded in large single blocks without putting undue pressure on their price. This characteristic is commonly referred to as market ‘depth’.

¹³ For example, funds supporting defined benefit pension schemes, insurance policies (life or property and casualty) or backing the issuance of guaranteed investment contracts (GIC).

¹⁴ FRS 17 becomes mandatory in the U.K. at the same time as the new International Accounting Standards (IAS) become mandatory for companies listed on European Union stock exchanges, that is for accounting periods ending on or after January 1, 2005. FRS 17 stipulates immediate recognition of gains and losses on a company pension scheme but in a secondary statement of gains and losses called ‘Total Recognised Gains and Losses’ rather than in the Profit and Loss account. IAS 19, the international standard relative to employee benefits has moved in the direction of FRS 17. Financial Accounting Statement No. 87 (FAS 87), the equivalent statement under U.S. generally accepted accounting principles (U.S. GAAP) issued in 1985 lags behind IAS 19 and FRS 17 both in the application of fair valuation and the rapid recognition of gains and losses.

recognised on the company balance sheet at their market value for assets or present value based on relevant gilt rates for liabilities.

III.A.1.5.3 Assessment

The assessment of market risk is now a very well-developed activity in the fund management industry. It has become part and parcel of performance assessment and, if not always done thoroughly *ex ante* by fund managers, it is certainly done *ex post* by a number of analysts in order to compare the so-called risk-adjusted performance of funds.

Ex-post analyses are usually pure statistical analyses of time series of returns. They produce estimates of return distributions. These estimates are then fed into risk-adjusted performance measures (RAPMs), the most common RAPM being the Sharpe ratio or ratio of expected excess return relative to the risk-free interest rate divided by the standard deviation of return, both on an annualised basis. There are some simple arguments why investors should prefer funds with the highest Sharpe ratios. However, Sharpe ratios may lead to unwarranted conclusions if applied to the comparison of funds with significantly different return distributions. For example, whereas a well-diversified traditional fund holding long security positions only may be expected to exhibit approximately log-normally distributed returns, a fund selling out-of-the-money options or implementing a dynamic strategy with similar consequences should exhibit a significant downward skewness and excess kurtosis of long-term returns. The Sharpe ratio would be inadequate to compare the performance of these two funds, but a generalised Sharpe ratio or some other RAPM accounting for skewness and kurtosis might do. See Chapter I.A.1 for a full discussion of RAPMs.

Because many traditional funds are limited in their choice of securities and/or investment strategies, they prefer to be judged on relative rather than absolute performance.¹⁵ Depending on their strategy they choose or create a benchmark and estimate their risk-adjusted performance relative to the benchmark. The standard deviation of returns relative to the benchmark is called the tracking error. The choice RAPM is the ratio of the average excess return relative to the benchmark over the tracking error; it is called the ‘information ratio’ or ‘appraisal ratio’. Often the analysis of performance is extended to a full performance attribution analysis to explain which strategies and which changes in market factors have contributed to profits and losses.

Ex-post assessments are certainly useful for comparison and analysis of returns as well as to check ex-ante assessments, but they do not replace the need for ex-ante assessments. Ex-post

¹⁵ Although, as Warren Buffet said, ‘You cannot eat a relative performance sandwich’, comparison to peers or to a chosen benchmark rather than absolute performance is seen as a clear indicator of skills and a powerful source of motivation.

assessments lack reliability and relevance because they are based on limited information – typically, a few years of monthly returns – and they are not forward-looking. Even if return data were available on a much more frequent basis, daily for instance, they could only lead to a more statistically accurate forecast of short-term returns. Methods such as exponential moving averages (EWMA) and GARCH (see Section III.A.3.4) have proved useful to estimate daily risks in financial markets exhibiting time-varying volatilities. But estimates of short-term volatilities have little relevance for long-term risks when these are governed by a specific investment strategy such as capital protection, and consequently short-term returns are not mutually independent.

It is only on the basis of ex-ante assessments of risks that fund managers can check and justify that they are adhering to their management mandate as described in a mutual fund prospectus or agreed with trustees or shareholders. Of course, ex-ante assessments are more difficult. One must rely on assumptions about future market behaviour and sometimes introduce a degree of subjectivity. One must also assume a trading strategy complete with limits and contingency plans. Too often ex-ante analyses are carried out using standard commercial models without sufficient questioning of the assumptions contained in these models. A typical error, for example, is to assume that future departures from the performance of a benchmark will be small if the tracking error has been small in the past. The problem is that, ex post, the tracking error is usually calculated on de-trended return series, but if the composition of the portfolio being evaluated is significantly different from the composition of the benchmark, the two return series may well have different trends.

III.A.1.5.4 Control/Mitigation

If star ratings from one to three were to mark the degree of difficulty of a task, the identification of market risks in fund management should be attributed only one star, risk assessment two stars and control/mitigation three stars. Control of limits is not much of a problem, but the design of a risk mitigation strategy is as complex as the design of the investment strategy itself. In fact, the two cannot be separated except in a few special circumstances.

III.A.1.5.4.1 Selective Hedging

As a first special case, some undesirable risks may have been acquired as part of a package and need to be reduced. A typical example is that of a fund investing in a particular industry sector worldwide but wishing to maintain currency exposures to a minimum. In this case forward currency contracts can be used as hedges, but these hedges will have to be readjusted as a function of changes in foreign-denominated asset values and rolled over regularly. The corresponding costs and residual uncertainties will need to be estimated.

A more complex case is that of positioning a bond portfolio to take advantage of some interest-rate changes whilst protecting the portfolio against other possible interest-rate changes. Active managers of bond portfolios seek to exploit specific views on interest rates, for example, that an interest-rate term structure will flatten or that rates in two currencies will converge. At the same time they are likely to want to reduce exposures to interest-rate movements that should not affect their strategies, for example, for the two strategies above, parallel shifts of all interest rates. A traditional method to achieve this is to calculate the first- and second-order derivatives of bond values with respect to their yield to maturity (see Section I.B.2.6). These derivatives or ‘sensitivities’ are called ‘value duration’ and ‘value convexity’ respectively.¹⁶ Assuming the same changes in yields across all bonds in a portfolio, the portfolio value duration and value convexity are simply obtained by adding up the individual bond value durations and value convexities. Adjusting the composition of a bond portfolio so that these two sensitivities become negligible is often interpreted as ‘immunising’ the value of the portfolio against parallel shifts in interest rates. An example of immunisation against a parallel shift of bond yields, hedging one bond with another, is given in Section I.B.2.7. But note that reducing the value duration and value convexity of a portfolio to zero does not eliminate all interest-rate risks. It is an efficient immunisation only against small parallel shifts in bond yields, which is a relatively unlikely scenario;¹⁷ many other variations of interest rates are possible.

A sensible approach to selecting a portfolio of bonds to be immune to some movements in interest rates whilst maximising the profit opportunity from a forecast movement is first to calculate each bond price variation relative to each relevant interest-rate movement and then to choose the portfolio weights so as to maximise the portfolio gain for the forecast interest-rate movement whilst leaving the portfolio value unchanged for the other movements.

III.A.1.5.4.2 Momentary Hedging

As a second special case, we have the momentarily undesirable risks. At times, fund managers may fear a correction in the markets that would harm their performance or may simply wish to reduce some exposures for peace of mind because they are momentarily absorbed by other tasks.

¹⁶ In the same way as we now say ‘value-at-risk’ rather than ‘dollar-at-risk’, it is time to say ‘value duration’ rather than ‘\$duration’. When minus the value duration is divided by the value of the bond, the result is called the ‘modified duration’. Finally, when the bond yield is expressed on an annual basis, ‘duration’ (or ‘Macaulay duration’) is defined as ‘modified duration’ multiplied by $(1 + \text{yield})$. Historically, Macaulay was the first author to introduce the concept of duration; he chose the name because, for a zero-coupon bond, it is equal to the maturity of the bond, and, for a coupon bond, it is equal to the average maturity of the cash flows weighted by their corresponding discount factors calculated at the bond yield. The second-order sensitivity relative of a bond value with respect to its yield is called ‘value convexity’ because it relates to the curvature of the value versus yield curve.

¹⁷ A parallel shift in bond yields corresponds approximately to a parallel shift in the zero-coupon rate curve. Actual movements of the zero-coupon rate curve are best captured by a principal component analysis (see Section III.A.3.7). The first principal component, which usually explains 75% to 80% of interest rates’ total variance, is frequently described as a parallel shift when in fact it is often anything but parallel. Medium-term rates (18 months to 3 years) are often more volatile than both short-term and long-term interest rates (see, for example, Alexander, 2001, Table 6.2b, p. 149).

Closing down some positions for a short while may prove difficult or expensive; an overlay hedge may be less costly and still efficient. Adding offsetting derivative positions does it. Even if the derivatives are not a perfect offset for the undesirable exposures, an approximate global hedge can be achieved. It may work particularly well during brief market crises when correlations between related market factors tend to increase. There are many examples throughout the Handbook of such strategies. For example, the use of call options on bond futures to hedge a bond portfolio is given in Example I.C.6.6. The use of equity swaps is explained in Section I.B.4.7.1.1.

III.A.1.5.4.3 Managing for a Risk-Adjusted Performance Target

Coming back to the general case, risk mitigation strategies are inseparable from investment strategies designed to achieve some risk-adjusted performance target. There is but one task for the active fund manager: to follow a policy – encompassing level of diversification, leverage, selection of securities, etc. – compatible with the objective of investors and his own forecasts. Like customers in a supermarket who want choice and want to know what they buy by reading the labels on the cans, investors want a description of the funds offered to them in terms of composition of assets, strategies, objectives and target risk levels. And fund managers must remain true to the description of their products. The type of assets in which a fund can invest is certainly a major determinant of the volatility of returns. For example, equities are generally regarded as more volatile than bonds; within equities some sectors such as dotcoms and emerging markets are clearly more volatile than, say, utilities in G7 countries, and so on. But other factors have also a large influence on risk; chief among them are the level of diversification and the degree of gearing of the risky assets. By increasing the number of relatively independent securities in a portfolio, diversification reduces total risk by averaging out the effect of specific, independent risks. Diversification can be optimised to obtain the best possible value of the chosen risk-adjusted performance target – Sharpe ratio, information ratio or other (see Section I.A.3.4 for details). Gearing up, or leveraging, means increasing the allocation of funds to a risky asset class relative to a risk-free asset class, usually cash deposits, or even borrowing in order to invest more in the risky assets than the equity value of the fund. The expected return above the risk-free rate and the volatility of return vary proportionally to the amount invested in the risky assets relative to the equity value of the fund. Therefore gearing up or down does not affect the Sharpe ratio of a fund (see Section I.A.3.5 for details), but it can be used to adjust the risk level to suit a specific group of investors. Of course, many investors are sophisticated enough to manage themselves their own gearing and diversification. All they want is to have a choice among a wide variety of well-defined funds.

This leaves fund managers with the choice of dynamic investment strategies as a means of controlling risk and return. Positions can be actively adjusted either with the objective of achieving a specific return distribution – we discuss a couple of examples in the following subsection – or simply to take advantage of evolving return forecasts. The latter is particularly difficult to optimise as the cumulative costs of rebalancing more frequently compared to the expected opportunity losses of rebalancing less frequently are difficult to perceive intuitively and to analyse quantitatively. This is a subject of academic interest (see Davis and Norman, 1990), but implementation of systematic dynamic strategies is lagging behind theory. Most active fund managers rely on heuristics to rebalance their portfolios. These are rules of thumb based on trial and error combining profit objectives with multiple limits: stop losses, delta limits, limits on turnover, etc. Only a small proportion of active fund managers – essentially those using highly quantitative investment strategies (e.g. cointegration or convertible arbitrage) or those promising protected returns – rely on systematic rebalancing rules.

III.A.1.5.4.4 Capital Protection

Since the early 1980s, there has been a growing number of funds offering some kind of performance protection in order to attract risk-averse investors. They try to offer the best of two worlds: on the upside a participation in the potentially high returns of a risky asset class – typically equities or commodities – or, as a minimum on the downside, a capital guarantee or the return on a low-risk investment – for example, a deposit or a bond.

With few exceptions capital guaranteed products have a stated maturity of a few years. Investors staying until maturity are guaranteed to receive a defined performance. For instance, I read the following offer received today: after 5 years you will be guaranteed 105% of the performance of the FTSE 100 index on your initial investment or your money back, whichever is the highest. The guarantee is from the sponsor of the product or a third party, usually a good-quality insurance company or bank. To add to the popularity of these products and attract small savers to long-term equity investments, tax advantages may also be available at maturity. On the other hand, early withdrawals are not guaranteed or are guaranteed at only a fraction of the initial investment.

Although these products may appear as manna from heaven to the unsophisticated investor, they are actually simple to manufacture. In our example, the sponsor could use the initial investment to buy the equivalent face value of a five-year zero-coupon bond and a FTSE 100, at-the-money, five-year over-the-counter (OTC) call option on 105% of the initial investment from a specialist bank.¹⁸ The zero-coupon bond might cost 75% and the call option 18%, leaving 7% to the

¹⁸ Always read the small print on the exact definition of the pay-off; it is often not as straightforward as it first appears.

sponsor to cover his expenses and contribute to profit. We leave to Section III.A.1.6.4 the manufacturing of the call option and, more generally, the dynamic hedging of option portfolios.

The main drawback of capital guaranteed investments is their bullet form: one fixed size, fixed maturity issue. Investors like the flexibility of open-ended funds whose shares can be issued or redeemed at any time at their net asset value plus or minus a small commission. Can any form of downside protection be offered on an open-ended fund? This question has exercised the minds of many financial engineers and only approximate answers have been found. In the early 1980s, many funds became interested in the concept of portfolio insurance and tried to implement it by themselves or with the help of consultants. Insurance of an equity portfolio consisted of overlaying short positions in the new equity index futures at critical times. A naïve strategy, for example, would be to short futures whenever the market index fell below a predefined level and to buy them back whenever the market index recovered above that level. This, in effect, is a very inefficient attempt at replicating a put option; it creates significant residual risks and costs, not to mention implementation difficulties (e.g. index jumps, lack of liquidity). Not surprisingly, during the crash of October 1987, this type of portfolio insurance disappointed and various dynamic portfolio strategies came under criticism (Dybvig, 1988).

Portfolio insurance strategies were improved (Black and Jones, 1987) and new concepts emerged, notably, that of constant proportional portfolio insurance (CPPI) (Black and Perold, 1992). Under CPPI a fund would maintain an exposure in a risky asset proportional to the net asset value of the fund above a certain minimum. For example, the risky asset could be an equity index future and the exposure would be 200% of the excess value of the fund above the value of 90% of the initial investment placed in a short-term money market. Thus the fund manager would promise (sometimes with a bank guarantee) as a minimum the money market return on 90% of the initial investment but would raise the expectation of an equity index performance with a leverage of up to 200%. In continuous markets and with frequent (weekly or daily) rebalancing, CPPI would be safe. In practice, negative jumps combined with a poor initial performance of the risky asset may bring the fund value rapidly to its minimum guaranteed level, at which point the fund becomes a pure money market fund. This is what has happened in the early 2000s to many CPPI funds that were launched at the end of the 1990s.

Special types of portfolio insurance strategies should be developed for funds that must meet specific liabilities. Future liabilities may be uncertain both in amounts and timing (e.g. insurance claims), nonetheless, fund managers must aim for a stable surplus of assets over liabilities. These issues are often obscured by ad hoc actuarial rules and regulations and are in great need of re-examination.

III.A.1.5.4.5 Compliance and Accountability

Investors are becoming increasingly sophisticated and capable of scrutinising the market practices and performance of fund managers. Beyond compliance with regulations and accounting standards, beyond the avoidance of market malpractice,¹⁹ fund managers must satisfy investors that they are following strategies compatible with stated performance objectives and risk limits. When results have been disappointing, investors have sued fund managers for negligence or non-compliance with agreed policies. The investor does not necessarily have to lose money for this approach to succeed. In a case that has set new standards of accountability for fund managers, the Chief Investment Officer of the Unilever Superannuation Fund (a pension fund) accused the Mercury Asset Management unit of Merrill Lynch Investment Managers of negligence after the fund had underperformed its benchmark by more than 10% over little more than a year (January 1997 to March 1998). Mercury had agreed a target of 1% per year above the benchmark return with a tracking error of no more than 3%. Although the return on the £1 billion fund had still been positive over the period, Unilever sought damages of £130 million. The case revolved about the improper use of risk assessment models and, crucially, about the delegation of day-to-day operations to a 'junior' investment officer. Merrill admitted no liability but, in June 2001, settled out of court for a substantial amount. Similar cases have followed since.

III.A.1.6 Market Risk Management in Banking

III.A.1.6.1 Market Risk in Banking

Banks, like hedge funds, take geared positions on various asset classes. They differ in that many of their assets (e.g. loans) are illiquid and some of their sources of funds (e.g. deposits on call) are low-cost but with an indeterminate term outside the banks' control. In addition, banks are engaged in a number of fee-earning activities, but that is not of primary interest as far as market risks are concerned.

Most market risks are taken by banks voluntarily with a view to benefiting from the exposures. At the same time, deposit taking from clients is based on trust and it is crucial for banks to maintain their reputation of financial stability and competence in managing risks. Otherwise, clients may slip away, causing funding to become rapidly more expensive, and the business may fall into a downward spiral.

But banks are generally well equipped to manage market risks. It is part of their core competences; they have powerful systems to analyse and monitor risks and good access to the

¹⁹ The fund management industry has recently been the subject of a series of enquiries about conflicts of interests due to close relationships between investment bankers and fund managers and about market malpractices such as 'market timing' which have resulted in hundreds of millions of dollars of fines, withdrawal of billions of dollars of funds, scores of firings, and the reorganisation of several financial conglomerates.

markets for hedging. Moreover, they operate under the close supervision of banking regulators.

III.A.1.6.2 Identification

As mentioned earlier (see Section III.A.1.2), a key distinction is made between liquid assets eligible for capital treatment under trading book regulations and less liquid assets or assets held with a long-term intent that are relegated to the banking book. Only assets in the trading book are subject to detailed statutory assessments and corresponding capital charges. By contrast, market risks in the banking book are largely ignored by banking supervisors, and so are market risks affecting liabilities. In fact, accrual accounting standards tend to hide such risks. Only if common sense indicates that unusually large market risks are present in the banking book will banking supervisors request some *ad hoc* estimates and monitoring/control procedures and be free to impose additional capital charges.²⁰

Within the trading book, market risks are traditionally categorised by main markets; so, interest-rate, equity, currency and commodity risks are often identified separately. Nonetheless many positions entail risks in several markets, for example, a share denominated in a foreign currency, a convertible bond, a commodity linked loan; such positions will have to be identified under each of the corresponding market risk types.

To facilitate market risk analyses, it is also traditional to distinguish primary from secondary risks and general market risks from specific risks. The primary risks are the directional risks resulting from taking a net long or short position in a given class of securities or commodities. The secondary risks are the other risks, deemed a priori to be less important, for example, volatility risk, risk on spread between a security and a futures contract on that security, a dividend risk, a risk on repo costs. Secondary risks may well appear less important than primary risks, except for the fact that many trading books are managed actively with the purpose of reducing primary risks to negligible proportions at the expense of an increase in secondary risks.

Likewise, general market risks, such as an exposure to movements of an equity index, may seem more important a priori than specific risks due to unequal variations of prices of shares in that index. However, the composition of a portfolio of shares may differ markedly and systematically from their weightings in a reference index, for instance because of lack of diversification or because of the implementation of a particular investment strategy (say, value strategy rather than growth strategy). Consequently, specific risks may be large compared to the general (also called 'systematic') market risk and should not be overlooked.

²⁰ Capital surcharges are left to the discretion of banking supervisors under Pillar II of the Basel Accord. See BCBS (2004b) for general principles on the management of interest-rate risks.

Finally, one should take into account the increasing proportion of option and option-like instruments in trading books. A financial option is an instrument offering the right but not the obligation for the owner to make a claim by a certain date if some underlying market factor (or combination of market factors) is favourable or, otherwise, to forgo any claim and gain nothing. Thus the pay-off of an option is a non-linear function of some market factors. By extension, instruments that yield a non-linear pay-off in some market factor(s) can be called option-like – for example, a bond price is a non-linear function of changes in the discount-rate curve. Because of this non-linearity, the fair value of an option or an option-like instrument depends on the full probability distribution of future values of the underlying market factor(s) and not only on their current or expected future values. Long-term volatilities and correlations affecting the value of options offer new trading opportunities and hence new market risks.

A systematic identification of market risks in the trading book should proceed through the identification/selection of key market factors and the construction of models relating the value of instruments in the trading book to these factors, a step that will be essential for the assessment of market risks.

Within the banking book, market risks are harder to identify because positions are generally not valued at fair prices and therefore fair price variations are of little concern. But it is common sense that most banks are exposed to large market risks in their banking books assets as well as on their liabilities. In fact, interest-rate maturity transformation, the short-term funding of longer-term loans, has been a traditional banking strategy since time immemorial and entails an exposure to interest-rate rises. The impact of interest-rate changes is also enhanced by the clients' behaviour: if there are any prepayment or extension possibilities on loans and deposits, clients will take maximum advantage of these options to make loans cheaper or deposits more attractive and therefore less profitable for the bank.

Some option-like positions in the banking book are particularly susceptible to interest-rate changes. For example, a line of credit at a predetermined spread above Libor is economically equivalent to an option on the credit spread of the client: the line is much more likely to be drawn upon when the creditworthiness of the client has declined than when it has been maintained.

III.A.1.6.3 Assessment

We stressed the multiplicity of market risk factors in the preceding section on identification: there are systematic and specific market risks; primary directional risks and secondary risks; volatilities and correlations. The assessment of market risks is based on the selection of a limited number of

market risk factors and a choice of models to describe uncertainties in the future values of these factors, their impact on the value of individual instruments and, consequently, on the values of portfolios.

Fortunately, a number of models have been put forward and tested over the last 30 years or so. They fall into three main categories: (i) probabilistic/statistical models describing uncertainties about the future values of market factors; (ii) pricing models relating the prices and sensitivities of instruments to underlying market factors; and (iii) risk aggregation models evaluating the corresponding uncertainties on the future values of portfolios of financial instruments. In the first category are the stochastic processes commonly used to describe the evolution of market factors: geometric Brownian motion, stochastic volatility models, GARCH models, etc. A prime example of the second type is the Black–Scholes; option pricing model (see Section I.A.8.7) which, for a given choice of dynamics for the underlying asset price, adds some efficient market assumptions and a hedging argument to yield a risk-free option price. The third type is exemplified by value-at-risk (VaR) models which, with the help of a few simplifying assumptions, produce a probability distribution (or at least some statistics) on the future value of a static portfolio at a chosen future time. These are explored in Chapters III.A.2 and III.A.3.

We should acknowledge immediately that, no matter how sophisticated mathematical models have become, they are idealisations of reality. They are bound to be approximate at best; at worst they can be misleading. Nonetheless, they are indispensable.

There is a large degree of subjectivity in the choice of a model. A balance must be struck between realism and tractability. Depending on the business at hand, different models may suit. Conversely, a particular choice of model is always vulnerable to ‘gaming’ by traders seeking to construct portfolios that will apparently exhibit little risk. That is why banking supervisors want to exercise some control over the use of models for regulatory risk reporting – they want to examine the way a model is used, by whom and for what purpose before ‘recognising’ its use for regulatory reporting and the determination of capital ratios.

The greatest degree of freedom in the choice of models lies at the very first stage in the choice of market factors and the description of their dynamics. For example, to describe interest-rate risks across portfolios of bonds, bond derivatives, swaps and other interest-rate derivatives, one starts with a choice of interest-rate term structure model. It can be one of many single- or multiple-factor models from which all interest rates are derived. The temptation on the front desk is to choose the simplest adequate model for each task at hand and therefore to use different models for different instruments. But a market risk manager should be concerned with

comprehensiveness and coherence across models so that the effects of a variety of possible interest-rate fluctuations are taken into account realistically and consistently.

Pricing models are indispensable for all securities that are not readily priced in the market, for example, most OTC derivative and structured products. They are also necessary for most securities whose prices are readily available in the market because, unless these prices are selected as market factors, we need to know how they would be affected by changes in the value of the selected market factors. For example, we need to know how the prices of bonds would be affected by some fluctuations in the risk-free interest-rate curve, such fluctuations being relative to the selected market factors. Pricing models should follow logically from the choice of dynamics in the underlying market factors; however, some simplifications/approximations are usually introduced to obtain realistic prices within a limited computation time.

Likewise, risk aggregation models should follow logically from the choice of market factors dynamics and pricing models. However, many further simplifications are introduced for two reasons:

- (i) Dependencies between market factors can be very complex; they may differ between normal and extreme market conditions, between short- and long-term horizons.
- (ii) Trading book portfolios are by definition very dynamic, but it would be complex to describe the effects of new business and dynamic hedging strategies.

The conventional wisdom in banks is therefore to concentrate on the short term under normal market conditions where dependencies may be approximated by linear correlations and portfolios may be assumed to remain relatively static. This is what regulators ask banks to do: to estimate the maximum level of market losses that would not be exceeded with a probability of more than 1% on a static portfolio over the following 10 trading days.²¹ This number is then back-tested against actual conditions and scaled up to produce a minimum capital requirement for market risks. Stress tests (explained in Chapter III.A.4) are then applied to ensure that the minimum capital requirements are sufficiently safe. Market risk capital requirements for portfolios assessed separately are then simply added together and added to capital requirements for other types of risks to yields the total minimum regulatory capital (MRC). Note that this naïve aggregation process is likely to produce a larger total MRC than necessary because it does not recognise the effects of diversification among risks. Unfortunately, however, it is not necessarily safe.²² In some cases, adding the VaRs may *understate* the gross risk.

²¹ This is what is commonly known as the VaR figure in banking.

²² The addition of VaR figures is not sub-additive in general. In particular, super-additivity may occur when the tail risks are bigger than if they were normally distributed.

Many banks have now taken a wider view of market risks than that requested by banking supervisors. Banks can choose parameters to suit their own internal purposes – whether to improve resource allocation, to set up an ideal level of capitalisation or to test strategic plans. They can choose their own time horizon for risk assessment and confidence level for extreme losses. They can incorporate less liquid instruments into the banking book, assume some initial pricing uncertainties and take into account the effects of trading and hedging strategies. In the banking book, they may develop models of customer behaviour in response to changes in interest rates and other market factors.

III.A.1.6.4 Control/Mitigation

Banks have the means and the competence to manage most market risks very effectively. First, they have some degree of control over the market risk they take. In the medium term they can shape the risks by modifying the design and pricing of the products they offer to their customers. They can also adjust their liabilities to a large extent to match the risk profiles of their assets. In the short term they can use derivative products to hedge most market risks if they wish to do so.

The importance of financial derivatives as market risk hedging instruments needs to be stressed. Financial derivatives have become a huge market. In terms of notional size of the underlying assets, they are twice as large as bond and equity markets combined, about \$140 trillion against \$70 trillion.²³ In terms of trading volumes they are even larger. However, the total market value of these instruments – counting the positive side only of each transaction – is less than \$3 trillion and, after netting exposures to single counterparties, less than \$2 trillion. The credit risk created by derivative products is therefore very small, about 40 times smaller than the credit risk created by bonds and equities. Derivatives are also relatively cheap to trade. As a fraction of notional size, the sum of bid–offer spread and commissions on derivatives is typically at least ten times cheaper than for the relevant underlying assets. Derivatives also exist on assets that would not be easy to trade, such as equity indices and notional bonds. These features make derivatives the choice instruments for hedging market risks,²⁴ and indeed banks are usually found on at least one side of most OTC derivative products and as active participants in listed derivatives markets.

The crucial element to set up a market risk control/mitigation strategy in a bank is the definition of the objective. As for fund managers and even more so, there are some undesirable market risks accumulated in the course of normal business; from time to time there may also be risks that should clearly be reduced because of changes in market or management circumstances. But

²³ Of the \$140 trillion total, about \$110 trillion is OTC and \$30 trillion listed; about 60% of financial derivatives are in the form of interest-rate swaps.

²⁴ Note that there are also financial derivatives to cover credit risks. The market for credit default swaps and other credit derivatives has grown at about 50% per year over the last 10 years and now covers about \$2.5 trillion of underlying assets.

the bulk of market risks taken by banks is still taken willingly with the objective of deriving a profit. The risk management objective must therefore be the optimisation of some risk-adjusted performance measure within the constraints on minimum regulatory capital and various concentration limits imposed by banking supervisors or adopted internally. This general objective is translated in the short term and at various hierarchical levels (division/desk/trader) into simpler objectives and limits. The simpler objectives usually also take the form of risk-adjusted performance measures, but with a cost of risk (or cost of risk capital) adapted to each management unit.²⁵

Having assessed market risks, recognised the tools that can be used for their control and defined the objective of market risk management, the design and implementation of a control/mitigation strategy should follow naturally. In reality, there are still some complexities due to people and organisations. First, individual incentives should be aligned with the stated objectives. Rewards cannot be based solely on results without considering and agreeing *ex ante* the risks being taken. When a market risk hedge is put in place, there is roughly one chance in two that the hedge will generate a loss. All too often, if the rationale for the hedge has not been clearly agreed at the start, a loss on the hedge will reflect badly on the hedger, especially if the risks being covered are risks that the bank used to accept in the past. Second, risk mitigation is achieved more economically at a macro than a micro level. The following case illustrates these two points.

In the late 1970s, the European markets became flooded with petro-dollars. International treasury divisions of banks grew rapidly to handle this ‘hot’ money that could flow in and out rapidly. Many international treasuries implemented a very cautious micro-hedging strategy: each dollar deposit had to be matched with a corresponding lending and vice versa, thus doubling the size of the balance sheet and losing a bid—offer spread to the market. At the same time in the same banks, domestic treasuries were continuing to run significant interest-rate gaps (longer interest-rate maturity schedules on the asset side than on the liability side) without worrying about it because they had always done so. Clearly, there was a lack of consistency between the risk management objectives and mitigation strategies between the domestic and international sides. It could be explained for a while by fear of the unknown on the international side. But eventually a more balanced approach had to be implemented in which the interest-rate maturity gap could be tracked on a net basis at regular intervals (e.g. daily) and managed globally. This achieved today in most banks and, indeed, international and domestic treasuries are now often

²⁵ Risks generated by various management units may have a different impact on global risk. Some may have a diversifying or even hedging effect; others may be highly correlated with global risk. The cost of risk attributed to each unit should therefore reflect the marginal contribution of each unit to global risk to ensure that local optimisations of risk-adjusted performance lead to a global optimisation of risk-adjusted performance. Decomposition of risk is explained more fully in Section III.A.3.6.

merged in a single treasury division implementing a coherent market risk management policy across desks.

We do not have the space here to detail market risk hedging strategies, but we can highlight a couple of points. First, managers are mostly concerned about reducing primary risks, that is, the risks associated with net long or short positions in various asset classes. Exposures to primary risks are characterised by first-order sensitivities, the ‘deltas’ to the corresponding market factors.²⁶ Delta is the word commonly used to describe the first-order sensitivity of or change in an option value per unit of underlying asset relative to a very small change in the underlying asset price, as explained in Section I.A.8.8. Primary risks can be covered (‘delta hedged’) relatively cheaply with futures and forward contracts.

In many instances, however, the exposures are not linear in the hedging instruments because of the presence of options or option-like instruments. Therefore, hedges with futures and forwards must be rebalanced over time as prices fluctuate. How often should delta hedges be rebalanced? As a rule of thumb, over a given time interval, the transaction costs of rebalancing a hedge (bid–offer spreads and commissions) increase as the square root of the rebalancing frequency, whereas the variance of residual risks decreases as the inverse of the rebalancing frequency. An optimum frequency (or more efficient rules based on actual market movements) can be derived from a choice of trade-off between costs and residual risk and the knowledge of some portfolio and market characteristics (volatility of underlying asset price, gamma or second-order sensitivity of the portfolio to changes in the underlying asset price, unit transaction costs); see, for example, Hodges and Neuberger (1989). The results may surprise traders because it is very difficult for anyone to gain an intuitive view about the right balance between expected costs and residual risks that accumulate slowly over time but can reach very large figures and can be very different from one portfolio to another.²⁷

Second, having more or less delta-hedged a portfolio, one is left with secondary risks now playing a primary role. Key among these risks are exposures to volatility changes and larger than expected underlying asset price movements, also known as gamma risks from the name generally

²⁶ When multiplied by the notional size of the underlying asset we obtain a so-called ‘dollar-delta’ or ‘delta-equivalent value’ that is the value of a position on the underlying asset having the same sensitivity as the option. Historically, other names have been used in other markets, for example ‘modified duration’ in the bond markets as we have seen in Section III.A.1.5.4.

²⁷ As a test, consider two portfolios A and B. A has twice the gamma of B (in dollar terms), twice the volatility and twice the transaction costs per unit volume. How frequently should B be rebalanced relative to A? The answer is, on average, four times more frequently than A. As a rule of thumb, the optimal rebalancing frequency is proportional to $(\text{volatility})^2 \times (\text{gamma}/\text{unit transaction cost})^{2/3}$, which is not immediately obvious.

given to second-order sensitivities to market factors.²⁸ The sensitivity of a portfolio to changes in volatility (assuming volatility can be described by a single parameter) is usually referred to as ‘vega’.²⁹ The two risks (explained in Section I.A.8.8) are related but are not the same. For a single plain vanilla option and using a constant volatility model, vega is equal to minus gamma multiplied by time to maturity. Obviously, for portfolios of options and option-like instruments with different maturities, there is no longer a simple relationship between vega and gamma. If the model for the dynamics of a market factor does not assume a constant volatility source of risk, then volatility may change over time and space (market prices) and one can no longer speak about a single vega, at least not without redefining what is meant by vega. Gamma risk is a concern inasmuch as if large and left unchecked it would require a very active and therefore expensive delta-hedging strategy and still leave the trader exposed to large residual risks in case of sudden market movements. Vega risk is a concern inasmuch as, for many market factors, volatilities may fluctuate rapidly (short-term volatilities may suddenly double or treble in a crisis) and are difficult to predict. Both gamma and vega risks can be controlled with the use of options. However, unless options for hedging can be found with maturities similar to the original exposures, vega hedging will be very crude and almost impossible to combine with gamma hedging. Hedging market risks remains therefore something of an art.

To summarise the degrees of difficulty in the identification, assessment and control/ mitigation of market risks in banking, I am minded to give the maximum three-star rating to all three tasks. At least this is my excuse for having given a longer description of these three tasks in the banking section compared to the fund management and non-financial firms sections.

III.A.1.7 Market Risk Management in Non-financial Firms

III.A.1.7.1 Market Risk in Non-financial Firms

Non-financial firms, whether in service, trading or manufacturing industries, take on market risks in the natural course of their business without seeking such risks to derive a profit. Their core competences lie elsewhere and they would rather unload these risks on to market professionals or hedge them directly in the markets. For example, in our global markets most manufacturers are exposed to foreign currency fluctuations; they affect the cost of raw materials, the price at which finished products can be sold in foreign markets as well as the price of competitive foreign imports.

²⁸ Because gamma is related to the curvature of the value of a portfolio as a function of a market factor, it is also referred to as ‘convexity’. That is certainly the case in bond markets when describing the second-order sensitivity of a bond price with respect to its yield.

²⁹ American traders, having quickly run out of Greek letters, opted for a hot blue star and an easy alliteration (i.e. vega and volatility).

Company reports are full of comments about business being affected by the weakness of one currency or the strength of another, by the cost of energy or raw materials, by the crippling effects of an interest-rate increase and the difficulties in raising capital. These are common market risks but they are often regarded by entrepreneurs as externalities about which they can do little. In fact more and more can be done to reduce these risks or at least smooth out their effects in the short to medium term. The real question is to what extent non-financial firms should design and implement hedging programmes to reduce the impact of market risks. Do such programmes add value to shareholders or are they simply contributing to the profits of banks and other financial intermediaries?

There are few guidelines on best practice for market risk management in non-financial firms and no regulations comparable to those in banking or fund management.

III.A.1.7.2 Identification

The identification of market risks in non-financial firms is arguably the most difficult of the three risk management tasks, followed by assessment and then control/mitigation. Thus, three stars for identification, two for assessment and only one for control/mitigation. Why? Because the management of financial risks is by definition not among the core competences of non-financial firms and therefore it tends to be neglected. It is a natural tendency that we tend to address the problems we know how to solve and ignore the others. So market risks may be not properly recognised, but if they were they might not be so difficult to evaluate and control.

In our modern, globalised and deregulated economies, market risks are very pervasive; they affect firms either directly or indirectly through competition. The three main sources of market risks are interest rates, foreign currency exchange rates and commodity prices. Equities tend to be the exception, except for holding companies and other companies relying heavily on investments in securities.

Finance directors sometimes ask: ‘What is less risky, borrowing at fixed rates or at floating rates?’ Here lies the paradox: on an accrual or cost-accounting basis, borrowing at fixed rates is safe, the financing costs are fixed, whereas floating rates are risky; but on a fair accounting basis it is the opposite, the present value of the floating rate debt is almost constant, whereas the present value of the fixed rate debt varies with interest rates like the price of the equivalent bond.³⁰ The choice

³⁰ If a loan is evaluated at a fair price like a bond and future cash flows are discounted at, say, the going Libor rates, it is easy to verify that the fair value of a properly priced floating rate note at Libor should be close to its face value at each interest-rate payment date. There may be small fluctuations between interest-rate payment dates and the present value of any credit spread, and profit margin above Libor will also fluctuate. The fair value of a fixed rate loan, on the other hand, will fluctuate like the fair value of the equivalent coupon bond, going down when interest rates go up and vice versa.

of accounting standards or, more generally, the choice of a coherent frame of reference for risk evaluation is critical. In particular, mixing the use of several frames of references can only lead to confusion. Difficult, subjective and inaccurate as it may be, I think that a fair valuation of assets and liabilities is the only acceptable basis for recognising and assessing risks, even though for other good reasons companies use accrual accounting extensively in their reports. We may need two sets of accounting principles: one to report results objectively and accurately, the other to serve as a rational basis for risk management.

But the answer to the previous question does not depend only on the choice of accounting standards, it also depends on the business, in particular on the composition of assets and liabilities. It is the uncertainty about the future equity value of the firm that is of concern to shareholders. Thus, if the assets of the firm are perceived to generate returns independent of future interest rates, a fixed rate funding may be the safer option, but if future returns are perceived as being highly correlated with interest rates then a floating rate funding is the safer option. In making these judgements, and considering the medium to long term, it may be helpful to consider inflation indices as intermediate factors. Future inflation rates are uncertain, but the operational profit margin of a business before financing costs can often be related to inflation indices and so are the financing costs. Real interest rates relative to inflation tend to be smaller and more stable than nominal interest rates. There is actually a growing market for inflation-linked bonds and loans that secure this relationship and thus are attractive to both investors and borrowers.

A similar approach should be used to recognise foreign exchange risk. Companies select a reporting currency, ideally the currency in which most assets and liabilities are denominated, most revenues and costs are incurred, and to which most shareholders are economically tied. When companies were mostly domestic, the choice was obvious. Now that many companies are truly international, there may be some doubt about the choice of a suitable reference currency. The easiest reference currency is the one with which most shareholders are comfortable, because the goal is to reduce risks for the shareholders. Thus if the majority of shareholders are based in the UK the preferred reporting currency should be the pound sterling, even if most of the business is conducted outside the UK; foreign exchange risks should be assessed on a sterling value and hedges against sterling should be put in place if the risks are deemed excessive.

Commodity and indirect market risks due to competition are often called economic risks or input/output risks rather than market risks, although many can be directly traced to market factors such as exchange rates rather than to non-market factors such as innovation, technology or regulation. They are terribly difficult to recognise and appreciate, but they are important.

Even purely domestically based companies may find themselves suddenly uncompetitive because of a flood of cheap imports brought about by the weakening of some foreign currency relative to their own domestic currency. It is difficult to appreciate these dangers in advance but critical to develop some awareness of them rather than ignoring them. An outsider's view may be informative. Risk managers can take their cues from equity analysts and rating agencies. They are experienced in detecting threats to individual companies and company sectors caused by possible changes in market conditions.

III.A.1.7.3 Assessment

To be approximately correct as a whole is more important than to seek accuracy in some areas and to ignore others. A key decision for assessing market risks in non-financial institutions is the choice of time horizon. For example, some companies assess foreign exchange risks purely on current payables and receivables, that is, to a horizon of perhaps three months. They argue that only those 'transactional' foreign exchange risks can be assessed accurately and hedged accordingly. A more fundamental question is whether the company is already exposed to exchange-rate fluctuations beyond this horizon because, for instance, it will not be able to adjust the price of its products and services within this time frame or it has long-term assets and liabilities denominated in foreign currencies. The latter has been called 'translation' or 'conversion' risk. Note that, unlike transaction risk, translation risk has no impact on cash flows so it is sometimes neglected. Similarly, longer-term transaction risks are sometimes ignored because they are less immediate, less certain and less precise. But an approximate evaluation of these further exposures is preferable to total ignorance. It will be a better basis for deciding not only what hedges to implement in the short term but also what offsets could be taken in the long term.

Long-term risk assessments extending to several years are indeed indispensable for deciding on major investments and developing long-term strategic plans. Companies face multiple choices that are risk-dependent such as where to locate a production facility, whether to outsource some services, whether to invest in new ventures with payback periods of many years, or whether to invest more now to maintain flexibility of choice at a later stage.

The assessment of long-term market risks and their potential impact on a firm therefore goes far beyond the calculation of VaR as carried out in banks. It calls for the application of decision analysis methods. A decision analysis cycle proceeds as follows. We construct a simple model of the objective under scrutiny (e.g. maximising the value of the firm) as a function of a few main market and other risk factors and for a base case strategy. Based on initial estimates of the risk factors, we calculate a base case value of the objective. Next we explore the sensitivity of the

base case to changes in initial estimates (typically we consider variations in a (subjectively) realistic range such as a 90% range) to identify the most significant sources of uncertainty as far as the objective is concerned. We also design alternative strategies that might do better depending on the evolution of the uncertain factors. At this stage we should understand what are the critical decisions and the most significant risk factors that would influence the choice of strategy. The following stage consists of introducing probability distributions to describe our state of uncertainty about the significant risk factors, and we deduce probability distributions for the objective value under alternative strategies. A choice of optimal strategy can then be attempted, taking into consideration the risk attitude of stakeholders in the firm. But among the possible choices there are often possibilities to acquire more information about some of the sources of uncertainty or to refine the basic model in order to determine the best strategy with greater accuracy and thereby to improve the objective. The decision analysis cycle should then be repeated with the updated information until no more economically valuable information or refinement can be found.

Note two essential points in this approach. First, the assessment of risks is carried out with the specific objective of improving decisions. Risk assessment is intimately combined with risk management. Second, market risk factors are combined with other sources of risks in this type of analysis. For decision-making, it does not matter what labels are put on risks; it is the combination of multiple risks and decisions – including responses from competitors – that is significant. The role of market risk specialists will be to alert management to the existence of certain risks and to contribute to the description of these risks. They cannot work in isolation; one risk is often contingent on another. A company may acquire a foreign exchange exposure if it wins a contract, but may not be sure to win, and even if it wins the foreign exchange exposure may vary as a function of fluctuating demand. These uncertainties and the corresponding decisions – pricing the bid, deciding on hedges, etc. – must be analysed simultaneously. Small firms may lack the expertise to carry out this type of analysis but there is no shortage of consultancy firms and financial intermediaries ready to help.

III.A.1.7.4 Control/Mitigation

The decision analysis method outlined in the previous section is particularly useful for making major strategic choices over the medium to long term. For example, a chemical company producing high-density polypropylene should consider whether naphtha or gas oil would be the more economical feed. A decision analysis may reveal that, due to uncertainties in the future costs of these two feeds, it is worthwhile to make the extra investment in a plant that can accept both feeds. That is a form of long-term market risk management. Likewise, a Japanese car

manufacturer producing cars for the US market may decide to locate production facilities in the USA to reduce foreign exchange risks rather than in a country with currently lower labour costs.

'Physical' long-term solutions limiting exposure to market risks are usually preferable to 'financial' hedges that could be considered as alternatives. For example, the chemical company could consider building a plant suitable for one feed and, in principle, purchase an OTC option on the excess cost of the second feed relative to the first. Likewise, the Japanese manufacturer could opt for the country offering the lowest production and delivery cost into the US market and hedge currency risks by entering into forward exchange contracts. But one should be aware of two likely problems with long-term financial hedges: liquidity and cash flow. Many financial derivatives markets are very deep, thus the Japanese manufacturer may find forward contracts in sufficient sizes to cover exchange-rate risks for the entire economic life of its plant; on the other hand, commodity derivatives are still relatively thin and it is very unlikely that the chemical company could find an OTC option to cover its risk over more than a few months.

The cash-flow problem is linked to liquidity. Many financial derivatives are liquid only over a relatively short term. When used to cover long-term exposures, positions in short-term derivatives are stacked up and rolled over. At every rollover, expiring contracts must be settled; between rollovers, margins must be posted. If unlucky, the hedger may accumulate large realised losses on the short-term contracts against unrealised gains on the initial exposure. The cash-flow problem thus created may prove fatal. The textbook case is MGRM, the US subsidiary of the German company Metallgesellschaft. In 1993 MGRM had accumulated positions on 154 million barrels of crude oil futures on the New York Mercantile Exchange (NYMEX) to hedge long-term supply contracts of crude oil at fixed prices it had agreed with its customers. Unfortunately for MGRM, crude oil prices started to decline and futures were in continuo (higher prices than spot), so that at each monthly rollover MGRM had to pay for the decline in prices of contracts it had bought a month earlier. By the end of the year the board of Metallgesellschaft decided that they could no longer afford to support the losses of their subsidiary. MGRM had made a simplistic calculation resulting in an over-hedge and misjudged the rollover risks and the risk of holding a very large proportion of the futures contracts (which led NYMEX to call for additional margins), but, most importantly, they had underestimated the potential cash-flow problem resulting from hedging 10-year exposures with short-term financial derivatives.

Thus, it is generally considered inappropriate to hedge translation risks or long-term transaction risks using derivatives. Translation risks relate to revaluation of foreign assets (such as subsidiaries) and liabilities rather than to cash flows. Traditional hedging strategies with derivatives would only be applicable if there is a plan to sell these assets or refinance liabilities in

a different currency, thus resulting in cash flows. Most firms prefer instead to hedge translation risks by matching assets and liabilities in the same currency. For example, a foreign subsidiary could be funded in the currency of that subsidiary rather than in the home currency. Hedging economic exposures and long-term transaction exposures with derivatives can also be problematic not only because of the mismatch between short-term and long-term cash flows but also because of the uncertainties attached to these future cash flows and the difficulty in predicting exactly how profits will be affected by a possible market movement. In such cases operational solutions such as those mentioned earlier can be preferable. Financial derivatives remain the choice instruments for hedging short-term transaction risks. If the right instruments are not available on exchanges, firms will find many banks willing to offer tailor-made OTC products.

But whether market risks in non-financial firms should be hedged at all remains an interesting question. Some argue against hedging as follows:

- (i) Short-term uncertainties will tend to average out naturally over the long term.
- (ii) Hedging is costly, and in the long term it only contributes to banks' profits.
- (iii) Shareholders and investors know what the risks are; these risks are already priced in the market or washed out by diversification.
- (iv) If competitors do not hedge a certain risk, it would be unsafe to be too much out of line: winning on the hedge might not be as favourable as losing would be damaging.

But others argue in favour of hedging that:

- (i) Market risks in non-financial companies serve no useful purpose. They are not chosen with the expectation of deriving a profit. Indeed, there is no risk premium in the pricing of those market risks that are diversifiable (e.g. currency risks).
- (ii) On the contrary, market risks create uncertainties in the performance of business units and the firm in general. This makes the planning process more difficult, obscures the true profitability of various activities and confuses the reward scheme.
- (iii) If unnecessary risks are eliminated, results become more stable, the debt–equity ratio can be increased and tax benefits can be reaped.³¹
- (iv) Reduction in risk can also make the firm more attractive to other stakeholders such as lenders, trade creditors, customers and employees (especially if they hold executive stock and have poorly diversified portfolios). Greater stability of earnings

³¹ Research into the use of derivatives by non-financial corporations suggests that derivatives are more likely to be used by firms with greater leverage. Or, vice versa, firms that reduce their markets risks can afford a higher leverage and a reduce cost of capital.

may mean that lower interest rates apply, trade terms are more advantageous, etc. This is known as reducing the costs of financial distress.

Different firms may reach different conclusions. Hedging market risks may be less important for large, diversified, internationally active firms than for smaller, more specialised firms. In large firms, market risks may already be well diversified and treasury departments may have the expertise to decide which risks are likely to be beneficial. In small firms, some market risks might be crippling and should be seen by most stakeholders as unnecessary, avoidable gambles, if only a proper hedging strategy were put in place.

III.A.1.8 Summary

There is a pervasive view today that market risk management consists essentially in calculating a value-at-risk. This introduction should help dispel this false impression. Instead, I hope the reader will have realised that market risks, although relatively well understood, are still hidden in many places, and any attempt to assess them is based on a large number of assumptions. And, of course, market risk management does not stop at the assessment phase but should lead to control and mitigation.

To start with the risk identification phase, one should not forget that there are market risks hidden in illiquid assets and liabilities that are not evaluated at fair value. It is not because a firm uses accrual accounting that these risks do not exist. That would be an ostrich-like, head-in-the-sand attitude. Fortunately, the wider adoption of the new International Accounting Standards favouring fair value and hedge accounting, the valuation of contingency claims (e.g. executive share option schemes) and the recognition of assets and liabilities heretofore not affecting reported company profits (e.g. company pension schemes) will help companies pay attention to market risks. A short-term effect of these changes may be to increase the volatility of company returns and make equity investments less attractive, but in the long term it will help better risk management and should result in a more efficient allocation of resources.

The risk assessment phase is mathematical but relies on the choice of an objective and coherent set of assumptions. The objective may be to assess the probability of insolvency within a year; this is what banking supervisors are focusing on. But it may also be to assess some risk-adjusted performance measure and improve resource allocation accordingly; or it could be any of a number of other objectives such as developing contingency plans. To each objective corresponds a reasonable set of assumptions, for instance, a choice of time horizon, whether normal or extreme market conditions should be considered, whether portfolios should be assumed to be

static or dynamic, whether the business should be regarded as a going concern or whether some assets should be valued on a fire-sale basis, and so on.

The risk control/mitigation phase follows logically from a choice of objective. Depending on the business, there may also be a number of constraints, regulatory or otherwise, limiting the level of acceptable market risk; some of these constraints may be biting. Nonetheless, there cannot be any logical control/mitigation strategy without a clear objective. Fortunately, the implementation of hedging and risk control strategies is now a lesser problem because of the existence of a deep, liquid and efficient market in financial derivatives. There are few market risks that cannot be adequately covered when there is a wish to do so. New hedging requirements create new derivatives markets, as we see happening with telecommunications bandwidths and pollution credits, to name just a couple of new commodity derivatives.

Market risk management is still a relatively new and growing field of expertise. To operate efficiently, the market risk management function must be independent of risk-taking functions as well as of the accounting and internal auditing functions, must be able to rely on adequate resources, must communicate regularly with risk-taking departments and senior management and, like internal audit, must have reporting lines through a general risk management function up to the board of directors. The quality of risk management directly affects risk-adjusted performance measures and, ultimately, shareholder value. As banking regulators remind us, ‘capital should not be regarded as a substitute for addressing fundamentally inadequate control or risk management processes’ (BCBS, 2004a, par. 723).

References

- Alexander, C (2001) *Market Models: A Guide to Financial Data Analysis*. Chichester: Wiley.
- BCBS (1996) ‘Amendment to the capital accord to incorporate market risks’ (January, modified September 1997). Available at <http://www.bis.org/publ/bcbs.htm>
- BCBS (2004a) ‘International convergence of capital measurements and capital standards’ (June). Available at <http://www.bis.org/publ/bcbs.htm>
- BCBS (2004b) ‘Principles for the management and supervision of interest rate risk’ (July). Available at <http://www.bis.org/publ/bcbs.htm>
- Black, F, and Jones, R (1987) ‘Simplifying portfolio insurance’, *Journal of Portfolio Management*, Fall, 48–51.
- Black, F, and Perold, A F (1992) ‘Theory of constant proportion portfolio insurance’, *Journal of Economic Dynamics and Control*, **16**, pp. 403–426.
- Davis, M H A, and Norman, A R (1990) ‘Portfolio selection with transactions costs’, *Mathematics of Operations Research*, **15**, pp. 676–713.
- Dybvig, P H (1988) ‘Inefficient dynamic portfolio strategies, or How to throw away a million dollars’, *Review of Financial Studies*, **1**, 67–88.

Hodges, S D and Neuberger, A (1989), Optimal replication of contingent claims under transactions costs,, *Review of Futures Markets*, **8**, pp. 222–239.

III.A.2 Introduction to Value at Risk Models

Kevin Dowd and David Rowe¹

III.A.2.1 Introduction

Value at risk (VaR) has been the subject of much criticism in recent years. Many of these criticisms relate to important precautions as to how VaR results should be interpreted as well as limitations on their use. Other criticisms, however, have been more sweeping, in some cases dismissing the entire concept as misdirected and wrong-headed. In that context, it is useful to consider how trading risk limits were determined before VaR became widely accepted.

Market risk arises from mismatched positions in a trading book that is marked to market daily based on uncertain movements in prices, rates, volatilities and other relevant market parameters. Market makers cannot operate successfully if they only broker exactly offsetting trades between customers. To be successful, they need to stand ready to execute trades on demand, and this inevitably results in open positions being created that are exposed to loss from adverse market movements. These open positions are hedged in the short run with less than perfect offsets. A common example is a dealer who executes an interest-rate swap with a customer in which he/she receives fixed and pays floating and then hedges by shorting government bonds and investing the proceeds in short-term instruments. There is still basis risk, since the spread between the swap and bond rates may change, but the major exposure to loss from a general rise in rates has been eliminated. In the longer term, the dealer will try to attract offsetting customer trades by shading future quotes to make such offsets attractive to the market. Failing that, the dealer may execute an offsetting swap with another dealer, although this is less desirable since it requires paying away a bid or offer spread instead of earning the spread on a customer deal.

Given that running a market-making function inevitably gives rise to market risk, institutions have always imposed restrictions on traders designed to limit the extent of such risk-taking. Until the early 1990s these limits were in the form of restrictions on:

- the size of net open positions, including delta equivalent exposures to movements in underlying rates and prices;
- the degree of maturity mismatch in the net position;
- the permissible amount of negative gamma in option positions;
- exposure to changes in volatility.

¹ Kevin Dowd is Professor of Financial Risk Management at Nottingham University Business School, UK, and David Rowe is Group Executive Vice President for Risk Management at SunGard Trading and Risk Systems in London.

These limits imposed a complex array of constraints on trader positions. They were difficult to enforce effectively, not least because risk exposures, such as those based on option gammas, often move quickly in response to market developments. The information and management systems of the time also meant that these limits were enforced piecemeal, with all sorts of inconsistencies and other undesirable results: ‘good’ risks were often passed over because they ran into arbitrary risk limits, decisions were made with inadequate appreciation of the risks involved, reducing risk in one area seldom allowed greater risk-taking elsewhere, and so forth.

Perhaps the most important shortcoming of this old system was the absence of integrated risk management. There was little coherence between the structure or management of the limits and the range of potential losses that they permitted to occur. Senior management committees charged with approving such limits were often at the mercy of technicians, and even the technicians were hard pressed to translate the limits into a consistent measure of risk, let alone provide an effective system of integrated risk management.

Gradually a consensus arose that what was of fundamental interest to the institution was the probability distribution of potential losses from traders’ positions, regardless of the exact structure of those positions. From this realization was born the concept of value at risk. This gave management a much more consistent way of embedding acceptable levels of risk into the formal limits within which traders were required to operate. Naturally, there was still a heavy dependence on market risk technicians to translate market dynamics and the traders’ positions into estimates of the risks being taken, but VaR did allow a firm’s management to define limits that reflected a well-considered risk appetite in a way that the pre-existing system did not. In that sense, one of the most important contributions of VaR has been an improvement in the quality of the management of risk at the firm-wide level.

III.A.2.2 Definition of VaR

VaR is an estimate of the loss from a fixed set of trading positions over a fixed time horizon that would be equalled or exceeded with a specified probability. Several details of this definition are worth emphasizing.

- *VaR is an estimate, not a uniquely defined value.* In particular, the value of any VaR estimate will depend on the stochastic process that is assumed to drive the random realisations of market data. The structure of the random process has to be identified and the specific parameters of that process must be calibrated. This requires us to resort to historical experience and raises a whole host of issues such as the length of the historical sample to be used and whether more recent events should be weighted more heavily than those further in the past. In essence, the goal is to arrive at the best

possible estimate of the stochastic process driving market data over the specific calendar period to which the VaR estimate applies. Moreover, it is also clear that market data are *not* generated by *stable* random processes. Differing methods for dealing with the uncertainty surrounding changes in these random processes are at the heart of why VaR estimates are not unique.

- *The trading positions under review are fixed for the period in question.* This raises difficult questions when the evaluation period is long enough to make this assumption unrealistic.² In this instance it is most common to scale up a VaR estimate for a shorter period on the assumption that market data move independently from day to day. Otherwise, it is necessary to model trades that mature within the specified time horizon and make behavioural assumptions relating to trading strategies during the period.
- *VaR does not address the distribution of potential losses on those rare occasions when the VaR estimate is exceeded.* It is never correct to refer to a VaR estimate as the ‘worst-case loss’. Analysis of the magnitude of rare but extreme losses must invoke alternative tools such as extreme value theory or simulations guided by historical worst-case market moves.

The use of VaR involves two arbitrarily chosen parameters – the *holding period* and the *confidence level*. The usual holding period is one day or one month, but institutions can also operate on other holding periods (e.g., one quarter or more), depending on their investment and/or reporting horizons. The holding period can also depend on the liquidity of the markets in which an institution operates. Other things being equal, the ideal holding period appropriate in any given market is the length of time it takes to ensure orderly liquidation of positions in that market. The holding period may also be specified by regulation. For example, Basel Accord capital adequacy rules stipulate that internal model estimates used to determine minimum regulatory capital for market risk must reflect a time horizon of two weeks (i.e. 10 business days). The choice of holding period can also depend on other factors:

- The assumption that the portfolio does not change over the holding period is more easily defended with a shorter holding period.
- A short holding period is preferable for model validation or backtesting purposes: reliable validation requires a large data set and a large data set requires a short holding period.

² One common example of this is the requirement to estimate VaR over a 10-day time horizon for purposes of calculating regulatory capital for market risk under the Basel Capital Accord.

The choice of confidence level depends mainly on the purpose to which our risk measures are being put. Thus, a very high confidence level, often as great as 99.97%, is appropriate if we are using risk measures to set capital requirements and wish to achieve a low probability of insolvency or a high credit rating. Indeed, the confidence levels required for these purposes can be higher than those needed to meet regulatory capital requirements. On the other hand, for backtesting and model validation, relatively lower confidence levels are desirable to get a reasonable proportion of excess-loss observations. For limit-setting, most institutions prefer confidence levels low enough that actual losses exceed the corresponding VaR estimate somewhere between two and twelve times per year (implying a daily VaR confidence level of 95% to 99%). This forces policy committees to take the size of the limit seriously, since losses over that limit can occur with a reasonable likelihood.

For the above reasons, among others, the ‘best’ choice for these parameters depends on the context. What is important is that the choices be clear in every context and be thoroughly understood throughout the institution so that limit-setting and other risk-related decisions are made in light of this common understanding.

III.A.2.3 Internal Models for Market Risk Capital

The original Basel Capital Accord was put into effect at the beginning of 1988. It set down rules for calculating minimum regulatory capital for banks based on a simple set of multipliers applied to credit-risky assets. The minimum capital calculation did not reflect risks associated with a bank’s mark-to-market trading activities, which were still quite small.

However, market risk factors were becoming more important constituents of the risk profile of most major money centre banks, and in the 1990s the Basel Accord was amended to reflect banks’ exposure to market risk. In a significant departure from traditional conventions, the new Amendment also allowed banks to employ their own internal VaR models to calculate their minimum regulatory capital for market risk. This permission was conditional on several requirements:

- The models and their surrounding technical and organisational infrastructure had to be reviewed and approved by the bank’s supervisor.
- The model used for calculating regulatory capital had to be the same one used for day-to-day internal risk management (the so-called ‘use test’).
- The VaR confidence level used in the regulatory capital calculation had to be 99%.
- The time horizon for the regulatory capital calculation had to be two weeks (i.e. 10 business days).

Assuming the above conditions were met, the capital requirement was to be at a level at least 3.0 times the 10-day VaR estimate averaged over the last 60 business days for any reporting period. Supervisors retained the prerogative of applying a larger multiplier if the results of backtesting exercises suggested that internal models were generating insufficiently high VaR estimates.

The extended holding period presented some questions. How should positions that matured during this time horizon be handled? What about trades likely to be booked to correct for the erosion of initial hedges due to ageing of the portfolio?

The amended Accord also offered banks a solution to many of these problems, which almost all of them adopted. This was to apply the ‘square root of time’ rule. That is, banks obtained 10-day VaR estimates by multiplying the daily VaRs by $\sqrt{10} \approx 3.16228$. This procedure effectively says that if traders took the same level of risk as indicated by the one-day VaR estimate for 10 consecutive days, the 10-day VaR estimate would be 3.16228 times the daily VaR. This rule is based on the formula for the distribution of the sum of random variables, assuming that daily returns are independent of each other (see Chapter III.A.3). Assuming a static portfolio for one day avoids most of the complications described above for longer time horizons. Moreover, in so far as even a one-day static portfolio is unrealistic, the consequences of this assumption will (hopefully!) be evident from the backtests on the VaR model that are described in Section III.A.2.8.

III.A.2.4 Analytical VaR Models

The assumption that holding period returns (i.e. b -day relative changes in value) are normally distributed provides us with a straightforward formula for value at risk. If our b -day returns R are normally distributed with mean μ and standard deviation σ , we write

$$R \sim N(\mu, \sigma^2) \tag{III.A.2.1}$$

as in Section II.E.4.4.1. Now if the portfolio is currently worth S , our b -day VaR at the confidence level $100(1 - \alpha)\%$ is given by

$$VaR_{b,\alpha} = -x_\alpha S \tag{III.A.2.2}$$

where x_α is the *lower α percentile* of the distribution $N(\mu, \sigma^2)$. That is, x_α is the number such that the probability that $R < x_\alpha = \alpha$ (see Section II.E.4.4.3). Since we require a fairly high degree of confidence, α is small (normally $0 < \alpha < 0.1$). Thus x_α will typically be negative. In fact, using the standard normal transformation (II.E.28) we can write $Z_\alpha = (x_\alpha - \mu)/\sigma$. In other words,

$$x_\alpha = Z_\alpha \sigma + \mu \tag{III.A.2.3}$$

where Z_α is the lower α percentile of the standard normal distribution. This can be obtained from standard statistical tables or from spreadsheet functions, such as the NORMSINV function in Excel. For instance, typing ‘=NORMSINV(0.05)’ into Excel gives the value -1.64485 for $Z_{0.05}$.

Putting together (III.A.2.2) and (III.A.2.3), we have derived the following simple analytic formula for VaR that is valid under assumption (III.A.2.1):

$$VaR_{h,\alpha} = -(Z_\alpha\sigma + \mu)S. \quad (\text{III.A.2.4})$$

Estimating VaR at a given probability using the normal distribution is very easy, once we have an estimate of the mean and standard deviation, as the following example shows.

Example III.A.2.1: Analytic VaR calculation

Suppose we are interested in the normal VaR at the 95% confidence level and a holding period of 1 day and we estimate μ and σ over this horizon to be 0.005 and 0.02, respectively. Now (III.A.2.4) tells us that for a portfolio worth \$1 million,

$$VaR_{1,0.05} = -(0.005 - 1.64485 \times 0.02) \times \$1 \text{ million} = \$27,897.$$

Note that the higher the confidence level, the greater the VaR. For instance, if we were interested in the corresponding VaR at the 99% confidence level, our VaR would be³

$$VaR_{1,0.01} = -(0.005 - 2.32634 \times 0.02) \times \$1 \text{ million} = \$41,527.$$

Applying the square root of time rule (explained in Chapter III.A.3), the corresponding VaRs over a 10-day holding period are:

$$VaR_{10,0.05} = \sqrt{10} VaR_{1,0.05} \approx 3.16228 \times \$27,897 = \$88,218,$$

$$VaR_{10,0.01} = \sqrt{10} VaR_{1,0.01} \approx 3.16228 \times \$41,527 = \$131,320.$$

Analytical approaches provide the simplest and most easily implemented methods to estimate VaR. They rely on parameter estimates based on market data histories that can be obtained from commercial suppliers or gathered internally as part of the daily mark-to-market process. For active markets, vendors such as RiskMetrics™ supply updated estimates of the volatility and correlation parameters themselves.

But while simple and practical as rough approximations, analytic VaR estimates also have shortcomings. Perhaps the most important of these is that many parametric VaR applications are based on the assumption that market data changes are normally distributed, and this assumption

³ ‘=NORMSINV(0.01)’ in Excel gives -2.32634 .

is seldom correct in practice. Assuming normality when our data are heavy-tailed can lead to major errors in our estimates of VaR. VaR will be underestimated at relatively high confidence levels and overestimated at relatively low confidence levels. Further discussion on this will be given in Chapter III.A.3.

But analytical approaches can also be unreliable for other reasons:

- Market value sensitivities often are not stable as market conditions change. Since VaR is often based on fairly rare, and hence fairly large, changes in market conditions, even modest instability of the value sensitivities can result in major distortions in the VaR estimate. Such distortions are magnified when options are a significant component of the positions being evaluated, since market value sensitivities are especially unstable in that situation.
- Analytic VaR is particularly inappropriate when there are discontinuous payoffs in the portfolio. This is typical of transactions like range floaters and certain types of barrier options.

In summary, analytic approaches provide a reasonable starting point for deriving VaR estimates, but should not be pushed too hard. They may be acceptable on a long-term basis if the risks involved are small relative to a firm's total capital or aggregate risk appetite, but as the magnitude of risk increases, and as positions become more complex, and especially more nonlinear, more sophisticated approaches are necessary to provide reliable VaR estimates.

III.A.2.5 Monte Carlo Simulation VaR

Fortunately, many problems that cannot be handled by analytical methods are quite amenable to simulation methods. For example, we might have stochastic processes that exhibit jumps or certain types of heavy tails that do not allow an analytical solution for our VaR, or the values of the instruments in our portfolio might be 'complicated' functions of otherwise straightforward risk factors, as is often the case with exotic options. Alternatively, our portfolio might be a collection of heterogeneous instruments, whose payoffs interact in ways that cannot be handled using analytical methods. And there again, we might have simple positions that can be handled using analytical methods, but are better handled using simulation. A good case in point would be a portfolio of long straddles: these options are simple, but their maximum loss occurs when there is no market movement at all. Although the VaR of such a position can be obtained using analytical methods, we have to be careful which analytical methods we apply. For example, delta-gamma methods, which are often used for options VaR, can be very treacherous when applied to such positions because they assume that the maximum loss occurs when underlying variables exhibit large moves. (These methods are discussed in Section III.A.2.7.6.) In such cases, we might prefer to use simulation methods, because we know they are reliable.

In these and similar circumstances, the most natural approach is to use Monte Carlo simulation, which is a very powerful method that is tailor-made for ‘complex’ or ‘difficult’ problems. The essence of this approach is first to define the problem – specify the random processes for the risk factors of the portfolio, the ways in which they affect our portfolio, and so forth – and then simulate a large number of possible outcomes based on these assumptions. Each simulation ‘trial’ leads to a possible profit/loss (P/L). If we simulate enough trials, we can then produce a simulated density for our P/L, and we can read off the VaR as a lower percentile of this density. In the following when the context is clear we drop the notation for the dependence of VaR on holding period b and confidence level $100(1-\alpha)\%$, writing simply ‘VaR’ for $VaR_{b,\alpha}$.

III.A.2.5.1 Methodology

To illustrate, suppose we wish to carry out a Monte Carlo analysis of a stock price S , and we assume that S follows a geometric Brownian motion process:

$$dS/S = \mu dt + \sigma dW \quad (\text{III.A.2.5})$$

where μ is its expected (per unit time) rate of return and σ is the spot volatility of the stock price. dW is known as a *Wiener process*, and can be written as $dW = \varphi(dt)^{1/2}$, where φ is a drawing from a standard normal distribution. Substituting for dW , we get

$$dS/S = \mu dt + \sigma\varphi(dt)^{1/2}.$$

This is the standard stock-price model used in quantitative finance. The (instantaneous) rate of change in the stock price dS/S evolves according to its drift term μdt and realisations from the random term φ . In practice, we would often work with this model in its discrete-form equivalent. If Δt is some small time increment, we approximate (III.A.2.5) by

$$\Delta S / S = \mu\Delta t + \sigma\varphi\sqrt{\Delta t} \quad (\text{III.A.2.6})$$

where ΔS is the change in the stock price over the time interval Δt , and $\Delta S/S$ is its (discretised) rate of change.

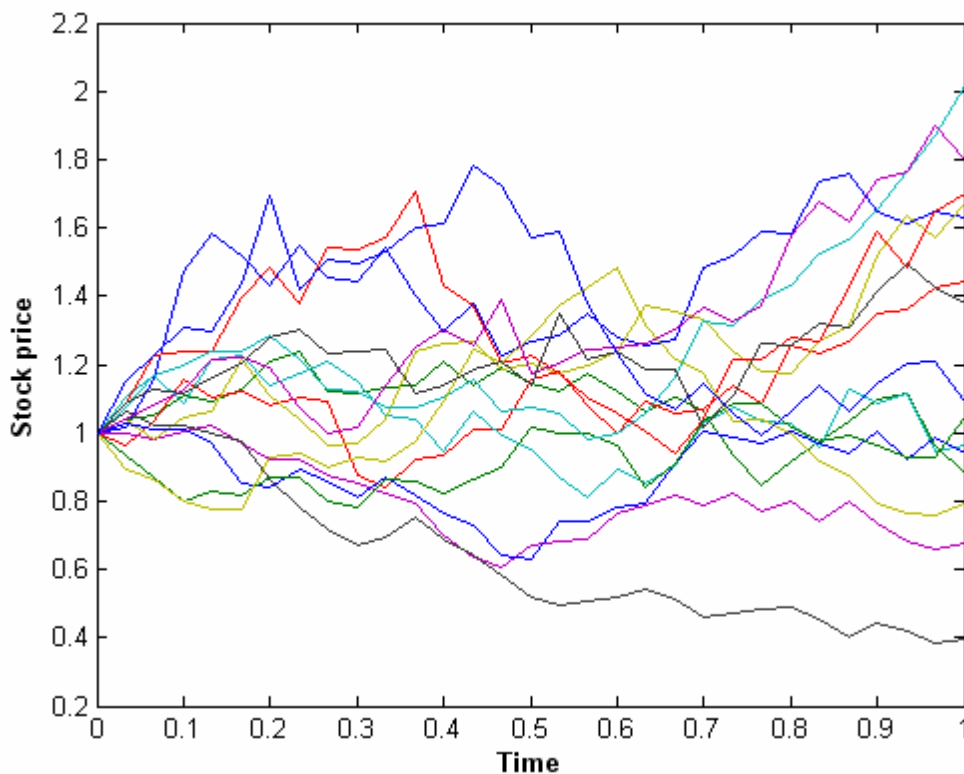
Note that (III.A.2.6) assumes that the rate of change of the stock price is *normally distributed* with mean $\mu\Delta t$ and standard deviation $\sigma\sqrt{\Delta t}$. Hence our criticisms of analytic VaR with respect to the normality assumption will also apply to the Monte Carlo VaR methodology, unless we employ an assumption for the underlying dynamics that is more appropriate than the geometric Brownian motions with constant volatility (III.A.2.5).

Now suppose that we wish to simulate the stock price over some period of length T . We would usually divide T into a large number N of small time increments Δt (i.e. we set $\Delta t = T/N$). We

take a starting value of S , say $S(0)$, and draw a random value of φ to update S using (III.A.2.6); this gives the change in the stock price over the first time increment, and we repeat the process again and again until we have changes in the stock price over all N increments. At this point, we have simulated the path of the stock price over the whole period T . We can then repeat the exercise many times and produce as many simulated price paths as we wish.

Some illustrative simulated price paths are shown in Figure III.A.2.1. We assume here that the starting value of our stock price, $S(0)$, equals 1, so each path starts from the 1 on the y -axis. Thereafter the paths typically diverge, moving randomly in accordance with their ‘laws of motion’ as given in the above equations. Moreover, since μ is assumed to be positive, there is a tendency for the stock prices to ‘drift’ upwards. The degree of dispersion of the simulated stock prices – the extent to which they move away from each other over time – is governed by the volatility σ . The bigger is σ , the more dispersed the stock prices will be at any point in the simulation. Note, too, that the simulated terminal stock prices will tend to approach the ‘true’ distribution of terminal stock prices as the number of draws grows larger. Even in this figure, which has only a limited number of paths, we can see that most of the terminal values are clustered around a central value, with relatively few in the tails. If we want to obtain a simulated terminal distribution which is close to the true distribution, all we need to do is carry out a large number of simulation trials. The larger the number of trials, the closer is the simulated terminal distribution to the true terminal distribution.

Figure III.A.2.1: Some simulated stock price paths



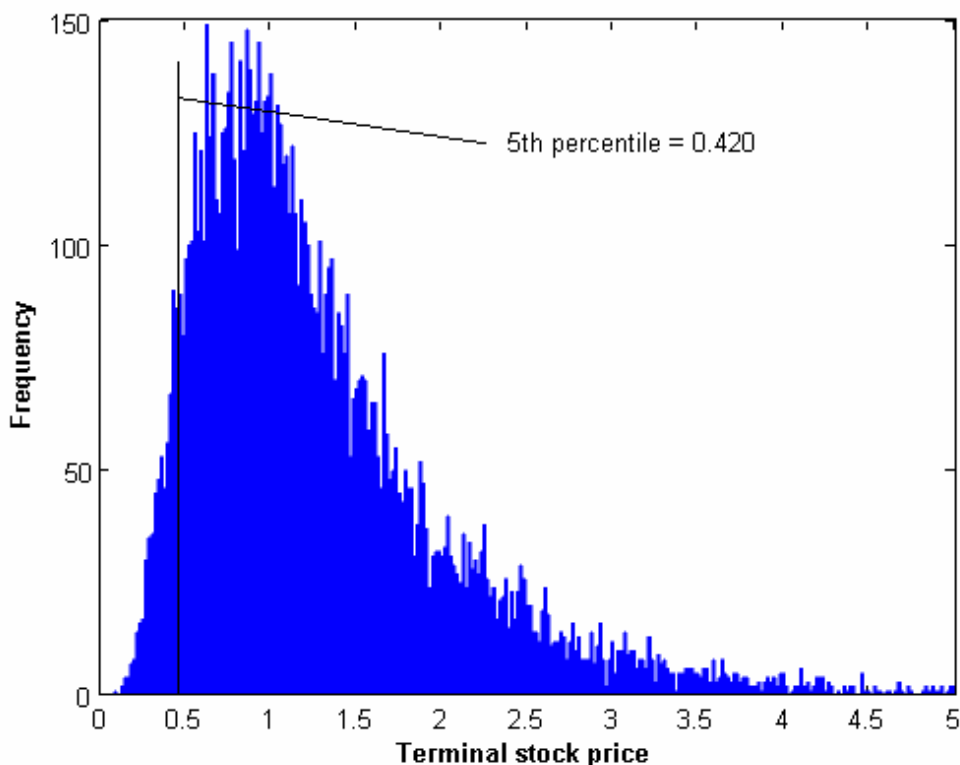
Note: Based on 15 Monte Carlo trials using parameters $\mu=0.05$, $\sigma=0.10$, $T=1$ and 30 step increments.

If we wish, we can estimate the VaR of the stock price by simulating a large number of terminal stock prices $S(T)$. We then read the VaR from the histogram of $S(T)$ values so generated. To illustrate, Figure III.A.2.2 shows the histogram of simulated $S(T)$ values from 10,000 simulation trials, using the same stock price parameters as in Figure III.A.2.1.

The shape of the histogram is close to a lognormal – which it should be, as the stock price is assumed to be lognormally distributed. The figure also shows the 5th percentile of the simulated stock price histogram. This percentile is equal to 0.420, indicating that there is a 5% probability that the initial stock price (of 1) could fall to 0.420 or less over the period, given the parameters assumed. A terminal stock price of 0.420 corresponds to a loss equal to $1 - 0.420 = 0.580$, so we can say that the VaR at the 95% confidence level is 0.580. This example illustrates how easy it is to estimate VaR using Monte Carlo simulation.

In addition, Monte Carlo simulation can easily handle problems with more than one random risk factor (see Section II.D.4.2).

Figure III.A.2.2: Histogram of simulated terminal stock prices



III.A.2.5.2 Applications of Monte Carlo simulation

Monte Carlo methods have many applications in market risk measurement and would be the preferred method in almost any ‘complex’ risk problem. Examples of such problems include the following, among many other possibilities:

- We might be dealing with underlying risk factors that are ‘badly behaved’ in some way (e.g. because they jump or show heavy tails) or we might have a mixture of heterogeneous risk factors. For example, we might have credit-related risk factors as well as normal market risk factors, and the credit risk factors cannot be modelled as normal.
- We might have a portfolio of options. In such cases, the value of the portfolio is a nonlinear (or otherwise difficult) function of underlying risk factors, and might be impossible to handle using analytical methods even if the risk factors are themselves ‘well behaved’.
- We might be dealing with instruments with complicated risk factors, such as mortgages, credit derivatives, and so forth.
- We might have a portfolio of heterogeneous instruments, the heterogeneity of which prevents us from applying an analytical approach. For example, our portfolio might be a collection of equities, bonds, foreign exchange options, and so forth.

III.A.2.5.3 Advantages and Disadvantages of Monte Carlo VaR

Monte Carlo simulation has many advantages over analytical approaches to calculating VaR:

- It can capture a wider range of market behaviour.
- It can deal effectively with nonlinear and path dependent payoffs, including the payoffs to very complicated financial instruments.
- It can capture risk that arises from scenarios that do not involve extreme market moves.
- Conversely, it can provide detailed insight into the impact of extreme scenarios that lie well out in the tails of the distributions, beyond the usual VaR cutoff.
- It lends itself easily to evaluating specific scenarios that are deemed worrisome based on geopolitical or other hard-to-quantify considerations.

The biggest drawbacks to the Monte Carlo approach to VaR estimation are that it is computer intensive and it requires great care to be sure all the details of the calculation are executed correctly. Non-technicians also find the process of imposing historically consistent characteristics on the scenarios to be quite impenetrable. Hence Monte Carlo VaR estimates are often viewed as coming from a black box whose credibility rests solely on the reputation of the technicians responsible for producing them. Nevertheless, Monte Carlo simulation is the most widely used approach to VaR estimation, and its popularity is likely to grow further as computers become more powerful and simulation software becomes more user-friendly. For large sophisticated trading operations, the only other widely used approach is historical simulation, to which we now turn.

III.A.2.6 Historical Simulation VaR

Historical simulation is a very different approach to VaR estimation. The idea here is that we estimate VaR without making strong assumptions about the distribution of returns. We try to let the data speak for themselves as much as possible and use the recent *empirical* return distribution – not some assumed theoretical distribution – to estimate our VaR. This type of approach is based on the underlying assumption that the near future will be sufficiently like the recent past that we can use the data from the recent past to estimate risks over the near future – and this assumption may or may not be valid in any given context.

III.A.2.6.1 The Basic Method

In applying basic historical simulation, we first construct a hypothetical P/L series for our *current* portfolio over a specified historical period. This requires a set of historical P/L or return observations on the positions currently held. These P/Ls or returns will be measured over a standard time interval (e.g. a day) and we want a reasonably large set of historical observations

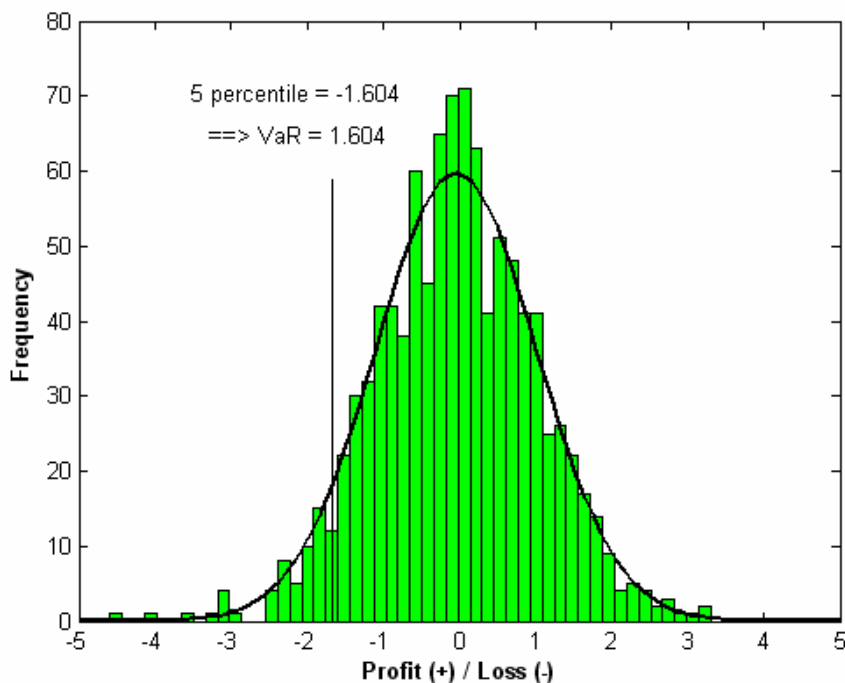
over the recent past. Suppose we have a portfolio of n assets, and for each asset i we have the observed return for each of T intervals in our historical sample period. (Our ‘portfolio’ could equally well include a collection of liabilities and/or instruments such as swaps, but we talk of assets for convenience.) If $r_{i,t}$ is the return on asset i in sub-period t , and if A_i is the amount currently invested in asset i , then the *simulated P/L* of our *current portfolio* in sub-period t is:

$$(P/L)_t = \sum_{i=1}^n A_i r_{i,t}.$$

Calculating this for all t gives us the hypothetical P/L for our current portfolio throughout our historical sample. This series will *not* be the same as the P/L *actually* earned on our portfolio in each of those periods because the portfolio actually held in each historical period will virtually never match our current positions.

Having obtained our hypothetical P/L data set, we can estimate VaR by plotting the data on a simple histogram and then reading off the appropriate percentile. To illustrate, suppose we have 1000 hypothetical daily observations in our P/L series (approximately four years of data for about 250 business days per year) and we plot the histogram shown in Figure III.A.2.3. If we take our VaR confidence level to be 95%, our VaR is given implicitly by the x -value that cuts off the bottom 5% of worst P/L outcomes from the rest of the distribution. In this particular case, this x -value (or the 5th percentile point of the P/L histogram) is -1.604 . The VaR at the 95% probability is the negative of this percentile, and is therefore 1.604.

Figure III.A.2.3: Historical simulation VaR



III.A.2.6.2 Weighted historical simulation

One of the most important features of basic historical simulation is the way it weights past observations. Our historical simulation P/L series is constructed in a way that gives any observation the *same* weight on P/L provided it is less than n periods old, and a zero weight if it is older than that. However, a problem with this is that it is hard to justify giving each observation in our sample period the same weight, regardless of age, market volatility, or anything else. For example, it is well known that natural gas prices are usually more volatile in the winter than in the summer, so a raw historical simulation approach that incorporates both summer and winter observations will tend to average the summer and winter P/L values together. As a result, treating all observations as having equal weight will tend to underestimate true risks in the winter, and overestimate them in the summer (see Shimko *et al.*, 1998). This weighting structure also creates the potential for ghost effects – we can have a VaR that is unduly high (low) because of a short period of high (low) volatility, and this VaR will continue to be high (low) until n days or so have passed and the observations have fallen out of the sample period. At that point, the VaR will fall (rise) again, but the fall (rise) in VaR is only a ghost effect created by the weighting structure and the length of sample period used. More detailed discussion of this point is left to Chapter III.A.3.

We can ameliorate these problems by suitably weighting our observations. In the natural gas case just considered, we might give the winter observations a higher weight than summer observations if we are estimating a winter VaR, and vice versa for a summer VaR.

Alternatively, we might believe that newer observations in our sample are more informative than older ones, and in this case we might age-weight our data so that older observations in our historical simulation sample have a smaller weight than more recent ones (see Boudoukh *et al.*, 1998).⁴ To implement an ‘age-weighted’ historical simulation, we begin by ordering our returns, worst return first. We then note the age of each return observation, and calculate a suitable age-related weight for each return. A good way to do so is to use an *exponentially weighted moving average* (EWMA). We choose a decay parameter λ , which indicates how much each observation’s weight decays from one day to the next. Again, a more detailed discussion of this methodology is left to Chapter III.A.3.

A worked example is provided in Table III.A.2.1 and in the Workbook entitled [Age-Weighted Historical Simulation](#). The first column gives the ordered returns, and the second gives the age of the corresponding return observation in days. For illustration only we assume an unrealistically

⁴ This approach takes account of the loss of information associated with older data and is easy to implement. However, it can aggravate the problem of limited data in the tails of the distribution.

small sample of only 150 observations. Basic historical simulation gives each return a weight of 0.00667, implying cumulative weights of 0.00667, 0.01333, 0.02000, etc. The historical simulation confidence level is 1 minus the cumulative weight plus 0.00667, and the VaR is the negative of the relevant observation. So, for example, in this case the historical simulation VaR at the 95% confidence level is 2.530% of the portfolio size, the point half way between the eighth and ninth worst simulated losses. However, if we apply exponential weighting using $\lambda = 0.97$, we get the weights given in the column headed ‘AW weight’. These are rather different from the historical simulation equal weights, and give more recent observations higher weights. The cumulative weights given in the next column are 0.02684, 0.04057, and so on. The rest of the analysis then proceeds as before. In this case, the EWMA weighted historical simulation VaR at the 95% confidence level is 2.659% of the portfolio value, falling between the fourth and fifth worst simulated losses. The effect of exponential weighting in this case is to raise the estimated VaR since some of the largest simulated losses occur relatively recently in the sample period.

Table III.A.2.1 also shows the same analysis conducted 25 days later. To illustrate the point, we have made the simplifying assumption that the positions are the same as on the first day so that all the simulated historical P/L values for any given calendar day are unchanged. We also assume that there are no large losses in the intervening 25 days, so that the worst simulated losses are the same as those recorded in the analysis at the initial date. Now, however, these observations have aged and their weights are lower, reflecting the observations’ greater age. Consequently, the cumulative weights are also lower and the impact is to increase the ‘effective’ confidence level for any given return observation. This, in turn, leads to a lower VaR for any given confidence level. In this case, the VaR at the 95% confidence level is 2.503%, a value interpolated between the ninth and tenth worst simulated losses. By contrast, the historical simulation VaR has remained unchanged because the historical simulation weights are unaltered.

Table III.A.2.1: Age-weighted historical simulation

Analysis at Initial Date

Ordered daily return	Age	Basic historical simulation				95% VaR
		HS weight	HS cum. weight	cl	VaR at chosen cl	
-3.50%	5	0.00667	0.00000	1.00000	3.500%	
-3.00%	27	0.00667	0.00667	0.99333	3.000%	
-2.80%	55	0.00667	0.01333	0.98667	2.800%	
-2.70%	65	0.00667	0.02000	0.98000	2.700%	
-2.65%	30	0.00667	0.02667	0.97333	2.650%	
-2.60%	50	0.00667	0.03333	0.96667	2.600%	
-2.57%	45	0.00667	0.04000	0.96000	2.570%	
-2.55%	10	0.00667	0.04667	0.95333	2.550%	
-2.51%	6	0.00667	0.05333	0.94667	2.510%	2.530%
-2.48%	17	0.00667	0.06000	0.94000	2.480%	
-2.45%	24	0.00667	0.06667	0.93333	2.450%	

Age-weighted historical simulation					95% VaR
AW weight	AW cum. weight	cl	VaR at chosen cl		
0.02684	0.00000	1.00000	3.500%		
0.01373	0.02684	0.97316	3.000%		
0.00585	0.04057	0.95943	2.800%		
0.00432	0.04642	0.95358	2.700%	2.659%	
0.01253	0.05074	0.94926	2.650%		
0.00681	0.06327	0.93673	2.600%		
0.00794	0.07008	0.92992	2.570%		
0.02305	0.07802	0.92198	2.550%		
0.02603	0.10107	0.89893	2.510%		
0.01862	0.12710	0.87290	2.480%		
0.01505	0.14572	0.85428	2.450%		

Analysis 25 Days Later

Ordered daily return	Age	Basic historical simulation				95% VaR
		HS weight	HS cum. weight	cl	VaR at chosen cl	
-3.50%	30	0.00667	0.00000	1.00000	3.500%	
-3.00%	52	0.00667	0.00667	0.99333	3.000%	
-2.80%	80	0.00667	0.01333	0.98667	2.800%	
-2.70%	90	0.00667	0.02000	0.98000	2.700%	
-2.65%	55	0.00667	0.02667	0.97333	2.650%	
-2.60%	75	0.00667	0.03333	0.96667	2.600%	
-2.57%	70	0.00667	0.04000	0.96000	2.570%	
-2.55%	35	0.00667	0.04667	0.95333	2.550%	2.530%
-2.51%	31	0.00667	0.05333	0.94667	2.510%	
-2.48%	42	0.00667	0.06000	0.94000	2.480%	
-2.45%	49	0.00667	0.06667	0.93333	2.450%	

Age-weighted historical simulation					95% VaR
AW weight	AW cum. weight	cl	VaR at chosen cl		
0.01253	0.00000	1.00000	3.500%		
0.00641	0.01253	0.98747	3.000%		
0.00273	0.01894	0.98106	2.800%		
0.00202	0.02168	0.97832	2.700%		
0.00585	0.02369	0.97631	2.650%		
0.00318	0.02954	0.97046	2.600%		
0.00371	0.03273	0.96727	2.570%		
0.01076	0.03643	0.96357	2.550%		
0.01216	0.04719	0.95281	2.510%	2.503%	
0.00870	0.05935	0.94065	2.480%		
0.00703	0.06805	0.93195	2.450%		

If we are concerned about changing volatilities, we can also weight our data by contemporaneous volatility estimates. The key idea – suggested by Hull and White (1998) – is to update return information to take account of changes in volatility. So, for example, if the current volatility in a market is 2% a day, and it was only 1% a day a month ago, then data a month old understate the changes we can expect to see tomorrow. On the other hand, if last month’s volatility was 1% a day but current volatility is 0.5% a day, month-old data will overstate the changes we can expect tomorrow.

Now suppose we are interested in forecasting VaR for day T . Again let $r_{i,t}$ be the historical return in asset i on day t in our historical sample, $\sigma_{i,t}$ be a forecast of the volatility of the return on asset i

for day t , made at the end of day $t - 1$, and $\sigma_{i,T}$ be our most recent forecast of the volatility of asset i . We then replace the returns in our data set, $r_{i,t}$, with volatility-adjusted returns, given by

$$r_{i,t}^* = \frac{\sigma_{i,T}}{\sigma_{i,t}} r_{i,t}.$$

Actual returns in any period t are therefore increased (decreased), depending on whether the current forecast of volatility is greater (smaller) than the estimated volatility for period t . We now calculate the historical simulation P/L as explained in Section III.A.2.6.1, but with

$r_{i,t}^*$ substituted in place of the original data set $r_{i,t}$.

The calculations involved are illustrated in Table III.A.2.2 and in the Workbook entitled [Vol-weighted Historical Simulation](#). Using the same returns as in Table III.A.2.1, this table shows two cases. In the first case, current daily volatility is 1.4%, generally above the range of contemporaneous daily volatilities over the historical dates with the worst losses. Most of the volatility weights are therefore greater than 1.0 and the volatility-adjusted changes in the portfolio are correspondingly greater (in absolute value) than the simulated historical changes. The confidence levels are the same as before. In this case, the effect of the volatility weighting is to increase the ‘effective’ changes and thus to increase the estimated VaR. In the second case, we have the same contemporaneous volatilities but a current volatility of only 0.8%. The volatility weights are now generally less than 1.0, and the ‘effective’ or volatility-adjusted changes are correspondingly reduced as is the estimated VaR. Note that the volatility weighting will generally alter the specific historical dates corresponding to the critical confidence level depending on the pattern of contemporaneous volatility. These dates are invariant, however, to the value of current volatility.

Table III.A.2.2: Volatility weighted vs. equal weighted historical simulation

Current volatility generally above contemporaneous historical volatility

Ordered Daily Return	Basic historical simulation			95% VaR
	Cumulative Weight	cl	VaR at Chosen cl	
-3.50%	0.00000	1.00000	3.50%	
-3.00%	0.00667	0.99333	3.00%	
-2.80%	0.01333	0.98667	2.80%	
-2.70%	0.02000	0.98000	2.70%	
-2.65%	0.02667	0.97333	2.65%	
-2.60%	0.03333	0.96667	2.60%	
-2.57%	0.04000	0.96000	2.57%	
-2.55%	0.04667	0.95333	2.55%	
-2.51%	0.05333	0.94667	2.51%	2.53%
-2.48%	0.06000	0.94000	2.48%	
-2.45%	0.06667	0.93333	2.45%	

Daily return	Vol-weighted historical simulation						95% VaR
	Contemp Volatility	Current vol	Vol weight	Ordered Vol-adjusted return	cl	VaR at Chosen cl	
-3.50%	0.82%	1.40%	1.7073	-5.98%	1.00000	5.98%	
-2.70%	0.80%	1.40%	1.7500	-4.73%	0.99333	4.73%	
-2.80%	0.90%	1.40%	1.5556	-4.36%	0.98667	4.36%	
-2.51%	0.85%	1.40%	1.6471	-4.13%	0.98000	4.13%	
-2.60%	0.95%	1.40%	1.4737	-3.83%	0.97333	3.83%	
-2.55%	1.00%	1.40%	1.4000	-3.57%	0.96667	3.57%	
-2.48%	1.05%	1.40%	1.3333	-3.31%	0.96000	3.31%	
-2.57%	1.10%	1.40%	1.2727	-3.27%	0.95333	3.27%	
-3.00%	1.30%	1.40%	1.0769	-3.23%	0.94667	3.23%	3.25%
-2.45%	1.25%	1.40%	1.1200	-2.74%	0.94000	2.74%	
-2.65%	1.50%	1.40%	0.9333	-2.47%	0.93333	2.47%	

Current volatility generally below contemporaneous historical volatility

Ordered Daily Return	Basic historical simulation			95% VaR
	Cumulative Weight	cl	VaR at Chosen cl	
-3.50%	0.00000	1.00000	3.50%	
-3.00%	0.00667	0.99333	3.00%	
-2.80%	0.01333	0.98667	2.80%	
-2.70%	0.02000	0.98000	2.70%	
-2.65%	0.02667	0.97333	2.65%	
-2.60%	0.03333	0.96667	2.60%	
-2.57%	0.04000	0.96000	2.57%	
-2.55%	0.04667	0.95333	2.55%	
-2.51%	0.05333	0.94667	2.51%	2.53%
-2.48%	0.06000	0.94000	2.48%	
-2.45%	0.06667	0.93333	2.45%	

Vol-weighted historical simulation							95% VaR
Daily return	Contemp Volatility	Current vol	Vol weight	Ordered Vol- adjusted return	cl	VaR at Chosen cl	
-3.50%	0.82%	0.80%	0.97561	-3.41%	1.00000	3.41%	
-2.70%	0.80%	0.80%	1.00000	-2.70%	0.99333	2.70%	
-2.80%	0.90%	0.80%	0.88889	-2.49%	0.98667	2.49%	
-2.51%	0.85%	0.80%	0.94118	-2.36%	0.98000	2.36%	
-2.60%	0.95%	0.80%	0.84211	-2.19%	0.97333	2.19%	
-2.55%	1.00%	0.80%	0.80000	-2.04%	0.96667	2.04%	
-2.48%	1.05%	0.80%	0.76190	-1.89%	0.96000	1.89%	
-2.57%	1.10%	0.80%	0.72727	-1.87%	0.95333	1.87%	
-3.00%	1.30%	0.80%	0.61538	-1.85%	0.94667	1.85%	
-2.45%	1.25%	0.80%	0.64000	-1.57%	0.94000	1.57%	
-2.65%	1.50%	0.80%	0.53333	-1.41%	0.93333	1.41%	

III.A.2.6.3 Advantages and Disadvantages of Historical Approaches

Historical simulation methods have both advantages and disadvantages. The advantages are:

- They are intuitive and conceptually simple, providing results that are easy to communicate to senior managers and interested outsiders (e.g. bank supervisors or rating agencies).
- Dramatic historical events (sometimes irreverently referred to as ‘the market’s greatest hits’) can be simulated and the results presented individually even when they pre-date the current historical sample. Thus the hypothetical impact of extreme market moves that are strongly remembered by senior management can remain permanently in the information presented, although not directly included in the VaR number.
- Historical simulation approaches are, in varying degrees, fairly easy to implement on a spreadsheet and can accommodate any type of position, including derivatives positions.
- They use data that are (often) readily available, either from public sources (e.g. Bloomberg) or from in-house data sets (e.g. collected as a by-product of marking positions to market).
- Since they do not depend on parametric assumptions about the behaviour of market variables, they can accommodate heavy tails, skewness, and any other non-normal features that can cause problems for parametric approaches, including Monte Carlo simulation.
- Historical simulation approaches can be modified to allow the influence of observations to be weighted (e.g. by season, age, or volatility).
- There is a widespread perception among risk practitioners that historical simulation works quite well empirically, although formal evidence on this issue is inevitably mixed.

The weaknesses of historical simulation stem from the fact that results are completely dependent on the data set. This can lead to a number of problems:

- If our data period was unusually quiet (or unusually volatile) and conditions have recently changed, historical simulation will tend to produce VaR estimates that are too low (high) for the risks we are actually facing.
- Historical simulation approaches have difficulty properly handling shifts that took place during our sample period. For example, if there is a permanent change in exchange-rate risk, it will take time for standard historical simulation VaR estimates to reflect the new conditions. Similarly, historical simulation approaches are sometimes slow to reflect major events, such as the increases in risk associated with sudden market turbulence.
- Most forms of historical simulation are subject to distortions from ghost effects stemming from updates of the historical sample.
- In general, historical simulation estimates of VaR make no allowance for plausible events that might occur but did not actually occur in our sample period.

There can also be problems associated with the length of our data period. We need a long data period to have a sample size large enough to get risk estimates of acceptable precision. Without this, VaR estimates will fluctuate over time so much that limit-setting and risk-budgeting becomes very difficult. On the other hand, a very long data period can also create its own problems:

- The longer the data set, the bigger the problem with aged data.
- The longer the sample period, the longer the period over which results will be distorted by past events that are unlikely to recur, and the longer we will have to wait for ghost effects to disappear.
- The longer the sample size, the more the news in current market observations is likely to be drowned out by older observations – and the less responsive will be our risk estimates to current market conditions.
- A long sample period can lead to data-collection problems. This is a particular concern with new or emerging market instruments, where long runs of historical data do not exist and are not necessarily easy to proxy.⁵

In practice, our main concerns are usually to obtain a long-enough run of historical data. As a broad rule of thumb, many practitioners point to the Basel Committee's recommendations for a minimum number of observations, requiring at least a year's worth of daily observations (i.e. 250

⁵ However, this problem is not unique to the historical simulation approach. Parametric approaches need a reasonable history if they are to use estimates (rather than just guesstimates or 'expert judgements') of the relevant parameters assumed to be driving the distributions. For historical simulation VaR it is sometimes possible to synthesize proxy data for markets prior to their existence based on their behaviour over a more recent sample period, but when doing so it is

observations, at 250 trading days to the year). However, such a small sample size is far too small to ensure that an historical simulation approach will give accurate and robust results. In addition, as the confidence level rises, with a fixed length sample, the historical simulation VaR estimator is effectively determined by fewer and fewer observations and therefore becomes increasingly sensitive over time to small numbers of observations. For example, at the Basel mandated confidence level of 99%, the historical simulation VaR estimator is determined by the most extreme two or three observations in a one-year sample, and this is hardly sufficient to give us a precise VaR estimate.

III.A.2.7 Mapping Positions to Risk Factors

Portfolio P/L is derived from the P/L of individual positions, and we have assumed up to now that we are able to model the latter directly. However, it is not always possible or even desirable to model each and every position in this manner. In practice, we project our positions onto a relatively small set of risk factors. This process of describing positions in terms of these standard risk factors is known as ‘risk factor mapping’. We engage in mapping for three reasons:

- We might not have enough historical data for some positions. For instance, we might have an emerging market security that has a very short track record or we might have a new type of over-the-counter instrument that has no track record at all. In such circumstances it may be necessary to map our security to some index and the over-the-counter instrument to a comparable instrument for which we do have sufficient data.
- The dimensionality of our covariance matrix of risk factors may become unworkably large. If we have n different instruments in our portfolio, we would need data on n separate volatilities, one for each instrument, plus data on $n(n - 1)/2$ correlations⁶ – a total of $n(n + 1)/2$ pieces of information. As n increases, the number of parameters that need to be estimated grows exponentially, and it becomes increasingly difficult to collect and process the data involved. This problem becomes particularly acute if we treat every individual asset as a separate risk factor. Perhaps the best response to this problem is to map each asset against a market risk factor along capital asset pricing model (CAPM) lines. So, for example, instead of dealing with each of n stocks in a stock portfolio as separate risk factors, we deal with a single stock market factor as represented by a stock market index.
- A third reason for mapping is that it can greatly reduce the necessary computer time to perform risk simulations. In effect, reducing a highly complex portfolio to a

extremely important to avoid overestimating the accuracy of the risk estimates by treating pseudo-data as equivalent to real data.

consolidated set of risk-equivalent positions in basic risk factors simplifies the problem, allowing simulations to be done faster and with only minimal loss of precision.

Naturally, there is a huge variety of different financial instruments, but the task of mapping them and estimating their VaRs can be simplified tremendously by recognising that most instruments can be decomposed into a small number of more basic, primitive instruments. Instead of trying to map and estimate VaR for each specific type of instrument, all we need to do is break down each instrument into its constituent building blocks – a process known as reverse engineering – to give us an equivalent portfolio of primitive instruments, which we can then map to a limited number of risk factors. There are four main types of basic building blocks. These are:

- spot foreign exchange positions;
- equity positions;
- zero-coupon bonds;
- futures/forward positions.

In this section we will examine the mapping challenges presented by each of these in turn, and compute the normal analytic VaR only, that is to say, we assume the risk factors to which positions are mapped have returns that are normally distributed. The calculation of VaR under more realistic assumptions for risk factor return distributions is discussed in the next chapter of the *Handbook*.

III.A.2.7.1 Mapping Spot Positions

The easiest of the building blocks corresponds to basic spot positions (e.g. holdings of foreign currency instruments whose value is fixed in terms of the foreign currency). These positions are particularly simple to handle where the currencies involved (i.e. our own and the foreign currency) are included as core currencies in our mapping system.⁷

We would then already have the exchange-rate volatilities and correlations that we require for the covariance matrix. If the value of our position is A in foreign currency units and the exchange rate (in units of domestic currency per unit of foreign currency) is X , the value of the position in

⁶ There are, of course, $n(n + 1)/2$ relevant values in an $n \times n$ symmetric correlation matrix, but we do not have to estimate the values on the main diagonal which are all identically equal to 1.0.

⁷ Where currencies are not included as core currencies, we need to proxy them by equivalents in terms of core currencies. Typically, non-core currencies would be either minor currencies or currencies that are closely tied to some other major currency (e.g. as the Dutch guilder was tied very closely to the German mark before introduction of the euro in both countries). Including closely related currencies as separate core instruments would lead to major collinearity problems; the variance–covariance matrix could fail to be positive definite, etc.

domestic currency units – or the mapped position – is AX . If we assume A to be a credit-riskless instrument bearing an overnight foreign interest rate, its value in units of foreign currency is constant and the only risk to a holder with a different base currency arises from fluctuations in X .

In this situation we can calculate VaR analytically in the usual way. For example, if the exchange rate is assumed to be normally distributed with zero mean and standard deviation σ_X over the period concerned, then

$$VaR = -Z_\alpha \sigma_X AX. \quad (\text{III.A.2.7})$$

The same approach also applies to other spot positions (e.g. in commodities), provided we have an estimate or proxy for the spot volatilities involved.

Example III.A.2.2: Foreign exchange VaR

Suppose we have a portfolio worth \$1 million, our ‘base’ currency is the pound, and $\pounds 0.65 = \$1$. This means that $X = 0.65$, $A = 1$ million (measured in dollars). We estimate the annual volatility of the sterling–dollar exchange rate to be 15%. The daily standard deviation is therefore $0.15/\sqrt{250} = 0.009487$ and the pound values of the daily VaRs at the 95% and 99% confidence levels are then given by

$$VaR_{1,0.05} = 1.64485 \times 0.009487 \times 0.65 \text{ million} = \pounds 10,143,$$

$$VaR_{1,0.01} = 2.32635 \times 0.009487 \times 0.65 \text{ million} = \pounds 14,346.$$

III.A.2.7.2 Mapping Equity Positions

The second type of primitive position is equity, and handling equity positions is slightly more involved. Imagine we hold an amount S_k invested in the common equity shares of firm k . If we treat every individual issue of common stock as a distinct risk factor, we can easily run into the problem of estimating a correlation matrix whose dimensions number in the tens of thousands. For example, if we wanted to evaluate the risk for an arbitrary equity portfolio drawn from a pool of 10,000 companies, the number of independent elements in the correlation matrix to be estimated would approach 50 million! It is no wonder that an alternative approach is desirable. In fact, a workable solution to this dilemma is the central concept in the CAPM, which was covered in detail in Chapter I.A.4.

The basic assumptions of the CAPM approach imply that the return to the equity of a specific firm k , R_k , is related to the equity market return, R_m , by the following condition:

$$R_k = \alpha_k + \beta_k R_m + \varepsilon_k$$

where α_k is a firm-specific constant, β_k is the market-specific component of firm k 's equity return and ε_k is a firm-specific random element assumed to be uncorrelated with the market. The variance of the firm's return is then:

$$\sigma_k^2 = \beta_k^2 \sigma_m^2 + \sigma_{k,s}^2$$

where σ_k^2 is the total variance of R_k , σ_m^2 is the variance of the market return R_m and $\sigma_{k,s}^2$ is the variance of the firm-specific random element ε_k for company k . The variance of the firm's return therefore consists of a market-based component $\beta_k^2 \sigma_m^2$ and a firm-specific component $\sigma_{k,s}^2$.

Assuming the firm's equity returns are normally distributed with zero mean, the VaR of an equity position currently valued at x_k in the shares of firm k is then:

$$VaR = -Z_\alpha \sigma_k x_k .$$

It is important to recognise that when we aggregate risk across many holdings in a well-diversified portfolio, the main contributor to the total will be the market-based component $\beta_k^2 \sigma_m^2$. Since the specific risk of each holding is assumed to be uncorrelated both to the market return and to all other specific risk elements, the share of total risk contributed by the specific risk terms falls continuously as the portfolio becomes more diversified and approaches zero when the portfolio approximates the composition of the total market.

Look again at the data requirements needed to be ready to estimate VaR based on an arbitrary portfolio drawn from a universe of 10,000 individual stocks. Instead of almost 50 million pairwise correlations we only need the market return volatility, 10,000 market betas for each stock and each stock's specific risk volatility (or more commonly every stock's total return volatility from which, with the market volatility and the individual stock betas, we can derive the specific volatilities). This is 20,001 parameters in all. Compared to a full correlation approach, the CAPM method requires a little more than 0.04% as many parameter values.

Estimating just the systematic risk of multi-asset equity portfolios using the CAPM approach reduces to a simple mapping exercise. Assume we have N separate equities holdings with market values of x_k for $k = 1, \dots, N$. Assume the betas of these holdings are β_k for $k = 1, \dots, N$ and the market return volatility is σ_m . Then the aggregate systematic VaR of the portfolio is:

$$VaR = -Z_\alpha \sigma_m \sum_{k=1}^n \beta_k x_k .$$

Thus, the systematic VaR is the appropriate critical value times the market volatility times a weighted sum of the market positions, where the weights are the respective betas for each holding. If we define X as the total market value of the portfolio we can modify the above equation slightly to read:

$$VaR = -Z_{\alpha} X \sigma_m \left[\sum_{k=1}^n (\beta_k x_k / X) \right]. \quad (\text{III.A.2.8})$$

In this form, the term in square brackets is the portfolio's *net beta*, or 'effective' market beta.

Example III.A.2.3: VaR of equity portfolio

Suppose we have a stock market portfolio with five stocks, labelled A , B , C , D and E . The value of the portfolio is \$1 million (so $X = \$1$ million) and we have equal investments in each stock (so $x_A/X = x_B/X = \dots = x_E/X = 0.20$). We are interested in a daily holding period, and the daily stock market standard deviation (σ_m) is estimated to be 0.025 (hence the annual volatility of the market is approximately 39.5%). The stock market betas are $\beta_A = 0.9$, $\beta_B = 0.7$, $\beta_C = 0.5$, $\beta_D = 0.3$, and $\beta_E = 0.1$. Substituting the relevant values into equation (III.A.2.8), the VaR at the 95% confidence level is equal to

$$VaR = -Z_{\alpha} X \sigma_m 0.2 \left[\sum_{k=1}^n \beta_k \right] = 1.64485 \times \$1,000,000 \times 0.025 \times 0.2 \times [0.9 + 0.7 + 0.5 + 0.3 + 0.1] = \$20,561.$$

There is no covariance matrix in the example above because all equities are mapped to the same single risk factor, namely, the market index, and this is highly convenient when dealing with equity portfolios. However, this single-factor mapping will underestimate VaR if we hold a relatively undiversified portfolio because it ignores the firm-specific risk, but the underestimation will often be fairly small unless the portfolio is very undiversified. We should also keep in mind that because it assumes a single dominating risk factor it can be unreliable when dealing with portfolios with multiple underlying risk factors (e.g. when there are significant industry concentrations within an equity portfolio).

III.A.2.7.3 Mapping Zero-Coupon Bonds

The third type of primitive instrument is a zero-coupon bond (often referred to simply as a 'zero' for short). We will assume for convenience that we are dealing with instruments that have no default risk. In this context it is standard practice to represent prevailing market conditions in terms of a continuously compounded zero-coupon interest-rate curve (sometimes also known as

a spot rate curve) across a selected set of future maturity dates. A simplified example of such a curve appears in Table III.A.2.3.⁸

Table III.A.2.3
Continuously Compounded Zero Coupon Discount Rates

Tenor	Today	3-Mos	1 Year	2 Year	3 Year	5 Year	7 Year
3-Mos	4.5000%	----->					
1 Year	5.0000%	----->	----->				
2 Year	6.0000%	----->	----->	----->			
3 Year	6.5000%	----->	----->	----->	----->		
5 Year	7.5000%	----->	----->	----->	----->	----->	
7 Year	8.0000%	----->	----->	----->	----->	----->	----->

In the context of Table III.A.2.3, we can have fixed cash flows maturing on any day out to seven years. Now consider the implications of this for mapping zero-coupon cash flows on arbitrary future dates to a set of fixed grid dates. Given the number of days to the defined grid dates at 3 months, 1, 2, 3, 5, and 7 years, we will interpolate linearly to obtain the effective zero rate on any arbitrary date within this grid.⁹ We now want to allocate (or map) a cash flow maturing on an intermediate date to the fixed grid dates in such a way that, to the maximum extent possible, the latter have the same risk characteristics as the original cash flow. What criterion will impose this effective risk equivalence?

A common approach is to require the present value impact of a one basis point (0.01%) change in the zero rate at the two surrounding grid points to be the same for the allocated cash flows as it was for the original cash flow. An example will help to illustrate how we can achieve this.

Example III.A.2.4: Mapping cash flows for a zero-coupon bond

Again using the simplified zero curve shown above, assume we have a risk-free cash flow of 1,000,000 maturing at 2.75 years. We want to create two cash flows, one at two years and one at three years, that have risk-equivalent characteristics to the one cash flow maturing at 2.75 years.

⁸ In practice, of course, we would have many more grid points at more frequent intervals than is shown in this simplified illustration.

⁹ In practice, in developed industrial countries, interest-rate futures are used to calibrate this curve to the market. We would use an overnight one-day rate as the basis to interpolate out to the next future roll date, which would be no more than three months in the future. It also should be noted that the practice of linear interpolation is not universally preferred because it implies there will be discontinuous changes in the slope of the zero curve at some grid points if the term structure is not uniformly linear over all horizons. To avoid this, some more complex form of interpolation such as cubic splines may be substituted. However, we will assume linear interpolation here to keep the example fairly simple.

We begin by recognising that the current interpolated zero rate for 2.75 years is:

$$r_{2.75} = r_2 \times (3 - 2.75) + r_3 \times (2.75 - 2) = 0.0600 \times 0.25 + 0.0650 \times 0.75 = 0.06375.$$

An increase of one basis point in the two-year rate would result in the 2.75-year rate rising to

$$0.0601 \times 0.25 + 0.0650 \times 0.75 = 0.063775.$$

Similarly, an increase of one basis point in the three-year rate would result in the 2.75-year rate rising to

$$0.600 \times 0.25 + 0.0651 \times 0.75 = 0.063825.$$

The ‘PV01’ (also called the present value of a basis point or PVBP) change in the two-year and three-year rates respectively will be the difference in the present value of the 1,000,000 cash flow after versus before these changes. Thus:

$$PV01_2 = 1,000,000(e^{-2.75 \times 0.063775} - e^{-2.75 \times 0.063750}) = 839,137.04 - 839,194.73 = -57.69,$$

$$PV01_3 = 1,000,000(e^{-2.75 \times 0.063825} - e^{-2.75 \times 0.063750}) = 839,021.67 - 839,194.73 = -173.06.$$

The next step is to find cash flows at two years and three years that have equal PV01 values. To do so, we want to solve the following two equations:

$$C_2 \times e^{-2 \times 0.0601} - C_2 \times e^{-2 \times 0.0600} = -57.69$$

and

$$C_3 \times e^{-3 \times 0.0651} - C_3 \times e^{-3 \times 0.0650} = -173.06.$$

We then rearrange slightly to get:

$$C_2 = -57.69 / (e^{-2 \times 0.0601} - e^{-2 \times 0.0600}) = -57.69 / -0.000377366 = 325,258.99$$

and

$$C_3 = -173.06 / (e^{-3 \times 0.0651} - e^{-3 \times 0.0650}) = -173.06 / -0.000246813 = 701,177.56.$$

Thus, the risk sensitivity to changes in the two-year and three-year zero rates of 1,000,000 at 2.75 years is the same as that of two cash flows of 325,258.99 at 2 years and 701,177.56 at 3 years. Put differently, the ‘original’ cash flow of 1,000,000 at 2.75 years ‘maps’ to these latter two cash flows, which can therefore be considered its mapped equivalent.¹⁰

¹⁰ Obviously this approach to mapping cash flows can be applied in the context of more complex methods for interpolation of the zero curve. In such cases the mechanics become more complicated, but the core principle of equivalent value sensitivity to small movements in the points that define the zero curve remains the same.

Example III.A.2.5: VaR of zero-coupon bond

Now suppose we wish to estimate the VaR of a mapped zero-coupon bond. To be more specific, we have a zero-coupon bond with 10 months to maturity, and the nearest reference horizons in our mapping system are three months and 1 year. This means that we need to map our single 10 months' horizon cash flow into (nearly) 'equivalent' cash flows at horizons of 3 and 12 months.

Now refer to the workbook, [VaR of a Mapped Zero Coupon Bond](#). Given 3-month and 12-month zero rates of 4.5% and 5%, we see that the interpolated 10-month zero rate is just under 4.9%. We then carry out the 'PV01' cash flow mapping and find that a 10-month zero with face value \$1m maps to (nearly) 'equivalent cash flows of \$719,217 at 3 months and \$654,189 at 12 months.

The second sheet of the book then performs the VaR calculation. Assume analysis of the historical data indicates estimated daily volatilities of 1.25% for the three-month rate and 1.00% for the 12-month rate. For the three-month rate this translates into an absolute volatility of $0.0125 \times 4.5\% = 0.0563\%$ or 5.63 basis points as a one standard deviation daily change. For the 12-month rate it translates into an absolute volatility of $0.01 \times 5.0\% = 0.05\%$ or 5.0 basis points as a one standard deviation daily change. Further assume an estimated correlation between the three-month and 12-month returns of 0.85. The previous worksheet showed that the cash flow mapped to the three-month maturity has a value sensitivity of \$17.78 to a one basis point change in the three-month rate, and that the cash flow mapped to the 12-month maturity has a value sensitivity of \$62.23 to a one basis point change in the 12-month rate. Based on this information about the value sensitivity to rate changes, the volatility of the rates and their correlation, the portfolio of mapped cash flows has a standard deviation of \$399.62.¹¹ Hence the VaR at the 95% confidence level (taking $Z_{0.05} = -1.645$) is $1.645 \times 399.62 = \$657.31$.

We can confirm this result by deriving the VaR directly from the standard deviation of the 10-month zero rate based on the standard deviations of the reference rates, their correlation and their relationship to the 10-month rate. The derived one-day standard deviation of the 10-month zero rate equals 4.995 basis points. Since the impact of a one basis point change on the value of the bond is \$80.003, the resulting VaR is $1.645 \times 4.995 \times 80.003 = 657.31$.

III.A.2.7.4 Mapping Forward/Futures Positions

The fourth building block is a forward/futures position. As explained in Chapter I.B.3, a forward contract is an agreement to buy a particular commodity or asset at a specified future date at a

¹¹ Using the standard formula for the variance of a portfolio (see Section I.A.2.3.2 and/or Section II.D.2.1), the variance is $399.62^2 = 159,696 = (17.779 \times 5.625)^2 + (62.2253 \times 5.000)^2 + 2 \times 0.85 \times 17.779 \times 5.625 \times 62.2253 \times 5.000$.

price agreed now, with the price being paid when the commodity/asset is delivered, and a futures contract is a standardised forward contract traded on an organised exchange. There are a number of differences between futures and forward contracts, but for our purposes here these differences are seldom important and we can treat the two contracts together and speak of a generic forward/futures contract.

To illustrate what is involved for VaR computation, suppose we have a forward/futures position that gives us a daily return that is dependent on the movement of the end-of-day forward/futures price. If we have x contracts each worth F , the value of our position is xF . If the return on each futures contract F is normal with standard deviation σ_F and zero mean return, the VaR of our position is

$$VaR = -Z_\alpha \sigma_F xF . \quad (III.A.2.9)$$

The main problem in practice is to obtain an estimate of the standard deviation σ_F for the horizon involved. Typically, we would have estimates for various horizons, and we would use some interpolation method to obtain estimates of interim standard deviations. Section I.C.7.5 presents some examples of VaR calculations for portfolios of commodity futures.

III.A.2.7.5 Mapping Complex Positions

Having set out our building blocks, we can now map more complex positions by producing ‘synthetic equivalents’ for them in terms of positions in our primitive building blocks.

(i) *Coupon-paying bonds.* We can map coupon bonds by regarding them as portfolios of zero-coupon cash flows and mapping each individual cash flow separately to its surrounding grid points. Of course, any one grid point can receive mapped cash flows from one or more actual cash-flow dates on either side of it. These multiple mapped cash flows, and more importantly their associated sensitivities, are aggregated into a single mapped cash flow and associated sensitivity for each grid point. The VaR of our coupon bond is then based on the VaR of its mapped equivalent in zero-coupon cash flows at the fixed grid points.

Example III.A.2.6: VaR of Coupon Bond

Suppose we have a coupon bond with an original maturity of 4 years and a remaining maturity of 3 years and 10 months. The workbook entitled [VaR of a Mapped Coupon Bond](#) illustrates the mapping process and associated VaR calculation. In this example we assume reference points for defining the yield curve at 3 months, 1, 2, 3, and 4 years. As with the zero-coupon bond, the observed daily volatilities for each rate are scaled by the rate levels to derive an absolute daily

standard deviation of the rate in basis points. We also assume a set of pairwise correlations of the daily changes in the rates. The correlations used in the example are arbitrary, but do reflect the tendency for changes in rates of similar maturities to be more highly correlated than those where maturities differ by a greater amount. In addition, pairs of rates with similar differences in maturities tend to be more highly correlated for rates at longer than at shorter maturities.

The worksheet then applies the standard matrix formula (II.D.4) for deriving aggregate volatility to estimate the VaR. Alternately, the mapped cash flows could be treated as the effective positions and analysed by either the Monte Carlo or historical simulation approach.

(ii) *Forward-rate agreements.* A forward-rate agreement (FRA) is an agreement to pay an agreed fixed rate of interest for a specific period starting at a known future date. It is equivalent to a portfolio that is long in a zero-coupon bond maturing at the end of the forward period and short in a zero-coupon bond maturing at the beginning of the forward period. We can therefore map an FRA and estimate its VaR by treating it as a long position in one zero-coupon bond combined with a short position in another zero-coupon bond with a shorter maturity.

Here there is one important distinction between FRAs and interest-rate futures. This is that interest-rate futures are settled at the beginning of the forward period, based on the discounted present value of the floating payment determined at that point. FRAs are settled at the end of the forward period at the undiscounted amount of the floating payment that was fixed at the beginning of the period. This means that there is some continuing market risk for an FRA up to the end of its forward period, whereas an interest-rate future is settled at the beginning of the forward period (adjusted for a short settlement delay) and thereafter has no impact on market risk.

(iii) *Floating-rate instruments.* Since a floating-rate note (FRN) reprices to par with every coupon payment, we can think of it as equivalent to a zero-coupon bond whose maturity is equal to the period until the next coupon payment. We can therefore map a floating-rate instrument by treating it as equivalent to a zero-coupon bond that pays its principal plus the current period interest amount on the next coupon date.

Example III.A.2.7: VaR of floating-rate note

The analysis of an FRN is very similar to that of a zero-coupon bond. The key point here is to appreciate that fixed-income theory tells us that we can price an FRN by treating it as a zero that pays the FRN's current coupon and principal at the FRN's next coupon date. Referring to the workbook [VaR of Mapped Floating Rate Note](#), we see that our FRN has a current coupon rate

of 5.5%. Given a face value of \$1,000,000, this means that we can treat it as a zero that pays \$1,027,500 at its FRN coupon date. Given that date is four months hence, this implies that our FRN maps to cash flows of \$1,212,992 in three months and \$39,406 in 12 months. Given earlier assumptions about rate volatilities and correlations, the portfolio of mapped cash flows has a standard deviation of \$185 and therefore a VaR of $1.645 \times \$185 = \304 .

(iv) *Vanilla interest-rate swaps*. A vanilla interest-rate swap (IRS) is equivalent to a portfolio that is long a fixed-coupon bond and short a floating-rate bond, or vice versa, and we already know how to map these instruments.

Example III.A.2.8: VaR of interest-rate swap

The workbook entitled [VaR of Mapped Vanilla Interest Rate Swap](#) illustrates how we deal with this type of instrument. To map an IRS, we first note that for our purposes the swap can be regarded as equivalent to the exchange of a coupon bond for an FRN: one leg of the swap has the cash flows of a coupon bond, and the other the cash flows of an FRN. However, for mapping purposes the FRN is equivalent to a zero, as we have just seen. Suppose, therefore, that we are the fixed-rate receiver on a vanilla IRS, and the fixed-rate leg of the swap has the same features as the coupon bond we have just considered (same notional principal, same term to maturity, etc.). Similarly, the floating-rate leg has the same features as the FRN we have just considered. Then it is ‘as if’ the cash flows from our swap are as given in Table III.A.2.4.

Table III.A.2.4: ‘As if’ cash flows from vanilla interest-rate swap

<i>t</i> (months)	4	10	16	22	28	34	40	46
Fixed rate leg	\$30K	\$30K	\$30K	\$30K	\$30K	\$30K	\$30K	\$1030K
Floating-rate leg	-\$1027.5K							
Net	-\$997.5K	\$30K	\$30K	\$30K	\$30K	\$30K	\$30K	\$1030K

These are not the ‘actual’ cash flows from the swap, but for mapping purposes we can consider them as if they were. In Table III.A.2.5 we map these ‘as if’ cash flows to obtain their (near) equivalents at the same reference points we used for the coupon bond, namely, 3 months, 1, 2, 3, and 4 years. (See the above referenced spreadsheet for the details of the mapping.)

Table III.A.2.5: Mapped cash flows from vanilla interest-rate swap

<i>t</i> (years)	0.25	1	2	3	4
Mapped Cash Flows	-\$1,156,000	\$16,140	\$59,662	\$258,256	\$843,749

We then derive an analytic estimate of the standard deviation of the portfolio’s daily change in value. We do this by scaling the value of a one basis point change in each reference rate times

that rate's one standard deviation change and applying the standard correlated aggregation procedure using the assumed correlations for the daily rate changes. The VaR is then the appropriate multiple of this standard deviation estimate. For the volatility and correlation assumptions made, the portfolio of mapped cash flows has a standard deviation of \$1719, and therefore a VaR (at the 95% confidence level) of $1.645 \times \$1719 = \2827 .

(v) *Structured notes*. These can be regarded as a combination of interest-rate swaps and conventional floating rate notes, which we can already map.

(vi) *Foreign exchange forwards*. A foreign exchange forward is the equivalent of a long position in a foreign currency zero-coupon bond and a short position in a domestic currency zero-coupon bond, or vice versa. Thus it will be sensitive to three market variables, the domestic interest rate, the foreign interest rate and the spot foreign exchange rate.

(vii) *Commodity, equity and foreign exchange swaps*. These can be broken down into some form of forward/futures contract on the one hand, and some other forward/futures contract or bond contract on the other.

III.A.2.7.6 Mapping Options: Delta and Delta-Gamma Approaches

The instruments just covered all have in common that their returns or P/L are linear or nearly linear functions of the underlying risk factors. Mapping is then fairly straightforward. However, once we have significant optionality in our positions, their values become nonlinear functions (often highly so) of the underlying risk factors. This nonlinearity can seriously undermine the accuracy of any standard (i.e. linear) mapping procedure. We should be wary of using linear-based mapping systems in the presence of significant optionality.

So how do we map option positions? The usual answer is to use a first- or second-order Taylor series approximation. We replace an option position with a surrogate position in an option's underlying variables and then use a first- or second-order approximation – often known as a delta or delta-gamma approximation – to estimate the VaR of the surrogate position. Such methods can be used to estimate the risks of any positions where the value is reasonably approximated by a quadratic function of its underlying risk factor(s). Thus, over a moderate range of variation, they can be applied to option positions that are nonlinear functions of an underlying cash price and to fixed-income instruments that are nonlinear functions of a bond yield. See Sections II.C.3.3, II.C.4 and II.D.2.3.

Suppose we have a simple European equity call option of value c . The value of this option depends on a variety of factors (the price of the underlying stock, the exercise price of the option, the volatility of the underlying stock price, etc.) but suppose we ignore all factors other than the underlying stock price, and use only the first-order Taylor series approximation of the change in the option value: $\Delta c \approx \delta \Delta S$ where Δc and ΔS are the changes in the option price and the stock price respectively, and δ is the option's delta.

If we are dealing with a very short holding period (so we can take δ as if it were approximately constant over that period), the option VaR is approximated by multiplying the underlying VaR by δ . Hence, if we further assume that S is normally distributed,

$$VaR \approx \delta Z_{\alpha} \sigma S \quad (III.A.2.10)$$

where S is the current price and σ is the standard deviation of returns over the relevant holding period. The new parameter introduced into the calculation, the option δ , is also readily available for any traded option and equation (III.A.2.10) can easily be extended to portfolios of options using the portfolio delta (see Section II.D.2.3.1). Hence the delta approach requires minimal additional data.

However, first-order approaches are only reliable when our portfolio is close to linear in the first place, and such methods can be very unreliable when positions have considerable optionality or other nonlinear features. If a first-order approximation is insufficiently accurate, we can try to accommodate nonlinearity by taking a second-order Taylor series (or delta-gamma) approximation:

$$\Delta c \approx \delta \Delta S + \frac{\gamma}{2} (\Delta S)^2.$$

Readers should refer to Section II.C.4.3 for the mathematical details of delta-gamma approximation and Section II.D.2.3 for further examples of its application.

For a long call position, both delta and gamma are positive. In this case, the gamma term contributes an increase in the value of the call option both when the stock price rises *and* when it falls. If the stock price rises ($\Delta S > 0$) the gamma term implies that the call option value rises by more than the delta-equivalent amount. If the stock price falls ($\Delta S < 0$), the call option value falls by less than the delta-equivalent amount. However, in the case of a short position in the call option, both delta and gamma are negative. Now if $\Delta S > 0$ the option value falls by more than the delta-equivalent amount and if $\Delta S < 0$ the option value will rise (become less negative) by a smaller amount than is implied by the delta term.

The bottom line is that *positive gamma contributes a favourable impact* to the value of a position, whether the price of the underlying rises or falls. Conversely, *negative gamma contributes an adverse impact* to the value of a position, whether the price of the underlying rises or falls. This observation motivates the following as one possible modification to the VaR for an option position to account for the second-order impact of the gamma term:

$$VaR \approx \delta Z_\alpha \sigma S - \frac{\gamma}{2} (Z_\alpha \sigma S)^2. \quad (\text{III.A.2.11})$$

Note that VaR is reduced if gamma is positive and increased if gamma is negative.

Example II.A.2.9: Option VaR (delta-gamma approximation)

Suppose we wish to estimate the VaR of a European call option. The parameters of the option are: S (underlying price) = X (strike price) = 1, r (risk-free rate) = 5%, option maturity = half a year, and σ (annual volatility of underlying) = 25%. We wish to estimate the VaR using a confidence level of 95% and a holding period of 10 days.

One approach is to use the delta approximation (III.A.2.10)

$$VaR \approx \delta Z_\alpha \sigma \sqrt{10/250} S$$

where the sigma term is multiplied by $\sqrt{10/250} = 1/5$ to convert the holding period from one year (250 business days) to two weeks (10 business days.) To make use of this approximation, we calculate the option delta using the standard formula for the Black–Scholes delta (see Table I.A.8.2). This gives us $\delta = 0.591$. We then input this value and the values of the other parameters into equation (III.A.2.10). This gives the delta approximation for the VaR of a call option as

$$0.591 \times 1.645 \times 0.25/5 = 0.0486.$$

If we wish to use a delta-gamma approximation instead, we obtain the option gamma using the standard formula for the Black–Scholes gamma (see Table I.A.8.2). This turns out to be 2.198. We then input the relevant parameter values into (III.A.2.11) to get

$$0.591 \times 1.645 \times 0.25/5 - (2.198/2) \times [1.645 \times (0.25/5)]^2 = 0.0412.$$

Taking account of the positive gamma term therefore reduces our VaR, as we would expect.

Now suppose we wish to do the same exercises for a long position in a European put. In this case, the option delta is -0.409 , and the gamma remains the same at 2.198. Applying equation (III.A.2.10) then gives us the delta approximation for the VaR as:

$$-(-0.409) \times 1.645 \times 0.25/5 = 0.0336.$$

The corresponding delta-gamma approximation is then found by applying equation (III.A.2.11):

$$-(-0.409) \times 1.645 \times 0.25/5 - (2.198/2) \times [1.645 \times (0.25/5)]^2 = 0.0262.$$

Again, taking account of the gamma term serves to reduce our VaR.

Much the same approach can be used to give us a second-order approximation to the VaR of a bond portfolio, with the first-order term reflecting the bond's duration and the second-order term reflecting its convexity.

III.A.2.8 Backtesting VaR Models

Backtesting involves after-the-fact analysis of the performance of risk estimation models and procedures. In the case of market risk there are two distinct types of backtesting that serve different purposes. These involve comparing ex-ante VaR estimates with ex-post values of (a) actual P/L in the applicable periods and (b) hypothetical P/L assuming positions remained static for the applicable period.

The first approach is the one required of banks under the market risk amendment to the Basel Capital Accord and can be thought of as an all-in test. Control of the actual recorded P/L is what risk systems are ultimately designed to achieve. Hence comparing VaR estimates to actual P/L must be a part of any backtesting process. However, one problem with such a test is that there is more than one reason why actual gains and losses may exceed the risk estimates unexpectedly often. This may occur because of weaknesses in the VaR estimation system, including (depending on the approach used):

- inaccurate historical market data and/or parameter estimation;
- incomplete consolidation of trading positions;
- inaccurate mapping of trades to risk-equivalent positions in a limited set of primitive securities;
- incorrect estimation of the portfolio standard deviation or excessive nonlinearities and/or non-normal return distributions resulting in inaccurate VaR estimates using the analytical approach;
- the generation of Monte Carlo scenarios that fail to match the target characteristics implicit in the estimated parameters;
- the use of inaccurate or insufficient historical time series to produce historical simulation VaR estimates.

It also may be caused by gains or losses from intra-day trading that are, by original design, not reflected in any of the three main approaches to VaR estimation. When backtesting reveals weaknesses in the VaR estimation system, it is important to be able to isolate the source of the problem. This is where the second form of backtesting is useful, although not always practiced.

The second approach involves comparison of VaR estimates with the hypothetical P/L that would have resulted in the applicable periods if all the trades at the beginning of each period were simply revalued based on end-of-period market prices. This eliminates the impact of day trading that affects the actual P/L, but one must be careful in constructing the hypothetical ex-post P/L results. Assume, for example, that these are based on the same valuation methods used in the VaR estimation process and that these methods are flawed because of one or more of the first three causes noted above. Then the comparison of VaR to these ex-post hypothetical P/L estimates may look acceptable when, in fact, they are both subject to the same flawed calculation methods. This can result in attributing inaccurate VaR estimates to the impact of day trading when the actual problem still lies in the estimation process itself.

When conducting the first and more traditional approach to backtesting, it is important to review the official P/L series to establish their relevance to the exercise. Accounting systems are not designed to maintain consistent time series except over fixed reporting periods such as a fiscal quarter. For shorter sub-periods such as a day, which are often the periods of interest for market risk estimation and backtesting, accounting systems attempt to maintain accurate period-to-date information only. Thus, if there is a mistaken entry that results in a large but erroneous daily gain or loss, this will be adjusted by a correcting entry the following day (or later) to bring the fiscal-period-to-date figures into line. Obviously this will result in misleading daily P/L figures for both the day the mistake was made and the day the correcting entry was booked. It is therefore important to compile the actual P/L to be used for backtesting on a current basis. This allows investigation and documentation of accounting errors and their appropriate treatment in the risk system while the details are still fresh in people's minds. Failing this, the actual P/L series may be sufficiently flawed to make meaningful comparison impossible.

Obviously the key comparison in backtesting is whether the actual or hypothetical P/L series exceeds the corresponding ex-ante VaR estimate (or, more generally, VaR estimates predicated on different confidence levels) with the predicted frequency (or frequencies). In most cases, actual P/L (adjusted for accounting errors and their subsequent correction) tends to produce a lower than expected frequency of observations outside the VaR estimate than is consistent with the probability used in those estimates. This appears to be because day trading allows positions to

be closed quickly when markets become volatile, thereby reducing actual losses compared to holding a static portfolio for a full 24 hours.

III.A.2.9 Why Financial Markets Are Not ‘Normal’

The *central limit theorem*, sometimes referred to as the *law of large numbers*, is a fundamental statistical insight with far-reaching consequences. It states that the distributions of sums and averages of random variables exhibit a traditional bell curve or normal distribution even when the individual variables are not normal. While exceptions exist, this holds true for almost any stable random variable found in nature. This gives the normal distribution certain plausibility, and normal distributions are in fact commonly observed in the natural sciences. Thus, it is not surprising that in the early days of modern finance there was some serious debate over whether the distribution of changes in market data departed from a normal distribution in a systematic way. Today the presence of high kurtosis or ‘heavy tails’ in such distributions is a well-accepted fact. In trying to incorporate such behaviour into market risk analysis, it is important to consider why such persistent departures from the pervasive normal distribution should occur.

A key assumption behind the central limit theorem is that the individual observations of random variables going into an average or sum are statistically independent. More often than not this is a reasonably good description of the thousands, or even millions, of individual buy and sell decisions that drive changes in demand and supply on any given day. Since the market clearing price reflects the net balance of these largely independent decisions, it is not surprising that changes in such prices often exhibit a roughly normal distribution. This is, however, not always the case.

Consider an example totally unrelated to finance. Suppose you equip the passengers of a single-deck cruise ship with a device that allows you to locate them exactly at any given moment. Then proceed to calculate once every minute the centre of gravity of all these locations with reference to the two-dimensional framework of the ship and plot the resulting distribution. At most times passengers will be in a variety of locations based on their personal preferences, their energy levels, their mood of the moment and the available alternatives. The resulting distribution of their centre of gravity over time will be a cloud of points bunched around the centre of the available passenger areas. We would expect it to exhibit something very close to a bivariate normal distribution.

Now assume there is an announcement over the ship’s loudspeaker that a pod of whales is breaching off the port bow. The consequences are fairly obvious. We would see a sudden outlier in the distribution as passengers rush to find a good viewing spot among the limited

spaces available. In the immediate aftermath of the announcement, all passengers know several things:

- There is an opportunity to see something quite unique.
- The time to see it is limited.
- There is an ideal location for viewing the phenomenon.
- *Everyone else knows what they know.*

It is this final point, this ‘mutual self-awareness’, that makes for the sudden mad rush to the port bow. Each passenger reacts to the knowledge that speed is of the essence if a good viewing place is to be secured. If the ship was nearly empty, or if only a few people were aware of the opportunity or were likely to take advantage of it (if, say, most passengers were confined to their cabins with sea sickness) the sense of urgency would be greatly reduced.

There is a relevant scene in the movie *Rogue Trader* about Nick Leeson and the Barings debacle. He is awakened by a call at home in the early hours of the morning from another member of the firm. The voice at the other end of the phone says urgently, ‘Turn on CNN!!!’ The TV in the bedroom flickers to life showing scenes of the Kobe earthquake. The voice at the other end of the phone says, ‘This is just going to *kill* the market!!!’

In effect this is much like the announcement on the ship but on a global basis. Observers around the world are suddenly focused on a common crystallising event with obviously directional implications for the market. In addition, *everyone knows that everyone else knows*. Suddenly the millions of decisions that drive the market are no longer randomly independent. Rather they are subject to a common shared perception. The core structural assumptions that underpin a normal distribution have temporarily broken down and we see a sudden extreme observation.

Various statistical methods are used to try to build such behaviour into distributions of risk factor returns. But what these approaches cannot do is predict in advance when such events will occur. Thoughtful consideration of such potential scenarios, especially those that present special threats given existing open positions in the book, is therefore an essential component of effective market risk management. Such analysis remains in the realm of experience and seasoned judgement that no amount of advanced analytical technique can replace. This topic will be discussed in detail in Chapter III.A.4.

III.A.2.10 Summary

In this chapter we have explained the underlying methodology for the three basic VaR model approaches – analytic, historical simulation and Monte Carlo simulation. A main focus of this

chapter has been the analytic VaR models that are only valid when portfolio values are linearly related to the underlying risk factors and portfolio returns are normally distributed. We have also considered simple analytic formulae for VaR based on delta-gamma approximation to simple portfolios of standard European options.

However, in reality things are not that straightforward. Portfolio returns are not normally distributed and, as is evident from Chapter I.B.9, options portfolios typically contain products with many underlying risk factors and various exotic features. The next chapter of the *Handbook* will consider how the basic VaR methodology that we have introduced here may be extended to more realistic assumptions about the products traded and the behaviour of asset returns.

References

Boudoukh, J, Richardson, M, and Whitelaw, R (1998) ‘The best of both worlds: a hybrid approach to calculating value at risk’, *Risk*, Vol. 11(5), pp. 64–67.

Hull, J, and White, A (1998) ‘Incorporating volatility updating into the historical simulation method for value-at-risk’, *Journal of Risk*, Vol. 1 (Fall), pp. 5–19.

Shimko, D B, Humphreys, B, and Pant, V (1998) ‘Hysterical simulation’, *Risk*, Vol. 11(6), p. 47.

III.A.3: Advanced Value at Risk Models

Carol Alexander and Elizabeth Sheedy^{1,2}

III.A.3.1 Introduction

The previous chapter introduced the three basic VaR models: the analytic model and two simulation models, one that is based on historical observations and another that uses a covariance matrix to generate correlated scenarios by Monte Carlo (MC) simulation. The fundamental assumptions applied in the basic forms of each model can be summarised as in Table III.A.3.1. A number of variations on these assumptions are in common use and some of these will be reviewed later in this chapter.

Table III.A.3.1: Basic VaR model assumptions

	Analytic VaR	Historical VaR	Monte Carlo VaR
<i>Risk Factor/Asset Distributions</i>	Normal	No Assumption	Normal
<i>P&L Distribution</i>	Analytic (Normal)	Empirical (Historical)	Empirical (Simulated)
<i>Requires Covariance Matrix?</i>	Yes	No	Yes
<i>Risk Factor/Asset Returns i.i.d.³</i>	Yes	Yes	Yes

Clearly the analytic and the MC VaR models are very similar. Indeed, if one were to apply the basic MC VaR model above to a linear portfolio – i.e. one without options, for which the portfolio value is a linear function of the underlying risk factors – the VaR estimate should be similar to the analytic VaR model estimate. If it were different, that would be because not enough simulations were used and a ‘small sample’ error had been introduced. But of course, one would not apply MC VaR to a linear portfolio; this method is quite computationally intensive, and would only be applied to portfolios containing options.

The Excel spreadsheet [SimpleVaR.xls](#) examines the VaR of a simple portfolio (\$1000 a point on the Johannesburg Top40 index, actually). From the single historical price series in the spreadsheet we compute the 1% 10-day VaR as:

- \$878,233 according to the analytic VaR model,

¹ Carol Alexander is Chair of Risk Management and Director of Research, ISMA Centre, Business School, University of Reading, UK. Elizabeth Sheedy is Associate Professor, Applied Finance Centre, Macquarie University, Australia.

² Many thanks to Kevin Dowd and David Rowe for their careful editing of this chapter.

³ i.i.d. stands for ‘independent and identically distributed’.

- \$839,829 according to the historical VaR model,
- something close to \$878,233 (hopefully) according to the MC VaR model.

The MC VaR model result changes every time we perform the simulation. In this spreadsheet only 1000 simulations are used so the result can vary a lot each time we press F9 to generate new simulations. However, if we used 100,000+ simulations, the MC VaR result would be \$878,233 (or very close to that) as in the analytic VaR model.

The main problem with the basic analytic and MC VaR models is that they are both based on the very restrictive assumptions that risk factor (or asset) returns are i.i.d. normally distributed. Most practitioners are aware that the assumption of normality is violated in reality, with returns often exhibiting skewness and leptokurtosis. This has led to the popularity of the historical simulation method.

The historical simulation method, which can be applied to any type of portfolio, can give results quite different from the other two methods, even for an extremely simple portfolio as shown above. If the portfolio P&Ls are non-normal, then the historical VaR should be more accurate, *assuming* the sample contains enough data points to estimate the 1% lower percentile of the historical distribution with sufficient accuracy. But this requires a very large amount of historical data, at least 1000 days.⁴ Even if the data were available, the portfolio price series is computed over these 1000 or so days using the *current* portfolio weights, and this may not be plausible (would *you* have traded the same portfolio 5 years ago, when the market was in all probability quite different?). And even if the current weights *are* plausible, we still have to assume that returns are i.i.d. and use the ‘square root of time’ rule to compute the 10-day VaR (see Section III.A.3.2).⁵

The use of a covariance matrix in the analytic and the MC VaR models has both advantages and limitations. One advantage is that the covariance matrix will not normally be based on a very large amount of historical data. Regulators, for some reason, require at least one year of historical data to be employed when computing regulatory capital using a VaR model (see Section III.A.3.4), but for internal purposes less than a year of data may be used as, for instance, in Section III.A.3.4.1. Hence the assumption of *current* portfolio weights is not as problematic as it is with the historical VaR model. However, an important limitation is that the use of a single covariance matrix assumes that all co-variations between risk factors are linear; but they are not

⁴ Regulators recommend 3–5 years of daily data, i.e. 750–1250 observations..

⁵ You cannot use 10-day returns instead, since these would have to be non-overlapping, and one would need a history spanning *decades* to obtain enough data!

linear. It is well known that the extreme variations in major risk factors can be more highly correlated than the ‘normal’ variations. We shall address this issue in Section III.A.3.5.3.

Thus, all three of the approaches are defective in different ways when the basic assumptions are used. The nub of the problem is this: how can VaR be estimated using assumptions that are consistent with the stylised facts we observe in financial markets? Most of this chapter is directed at this problem, which is fundamentally concerned with model risk. Section III.A.3.2 examines the standard distributional assumptions more closely and the ways in which they are violated. It concludes, somewhat counter-intuitively, that the most pressing problem for those modelling market risks is not heavy tails, but volatility clustering. Accordingly, Section III.A.3.3 examines two different approaches to modelling volatility clustering: EWMA and GARCH. These models are then applied to the problem of VaR estimation in Section III.A.3.4. Section III.A.3.5 examines some other solutions to the problem of heavy tails in VaR estimation: Student’s t , EVT and normal mixtures. The remaining sections of the chapter address some other advanced topics in VaR modelling. It is often desirable to decompose VaR into components for purposes of limit setting and performance measurement. These decomposition techniques are considered in Section III.A.3.6. Section III.A.3.7 examines principal component analysis, an important tool for analysing bond and futures portfolios. Section III.A.3.8 concludes the chapter.

III.A.3.2 Standard Distributional Assumptions

When measuring risk, for example in a VaR calculation, it is often necessary to make some assumption about the distribution of portfolio returns. The most widespread choice is the assumption that returns are i.i.d. normal. That is, each return is an independent realisation from the same normal distribution (see Section II.E.4.4). Normality implies that the distribution can be completely described with only two parameters: the mean and the variance.

Yet many financial analysts are sceptical about the assumption of normality. The skewness and (excess) kurtosis should equal zero, but this is rarely the case. To illustrate this point, let us consider daily returns for the USD/JPY for the period from January 1996 to July 2004 as illustrated in Figure III.A.3.1 and described in Table III.A.3.2.

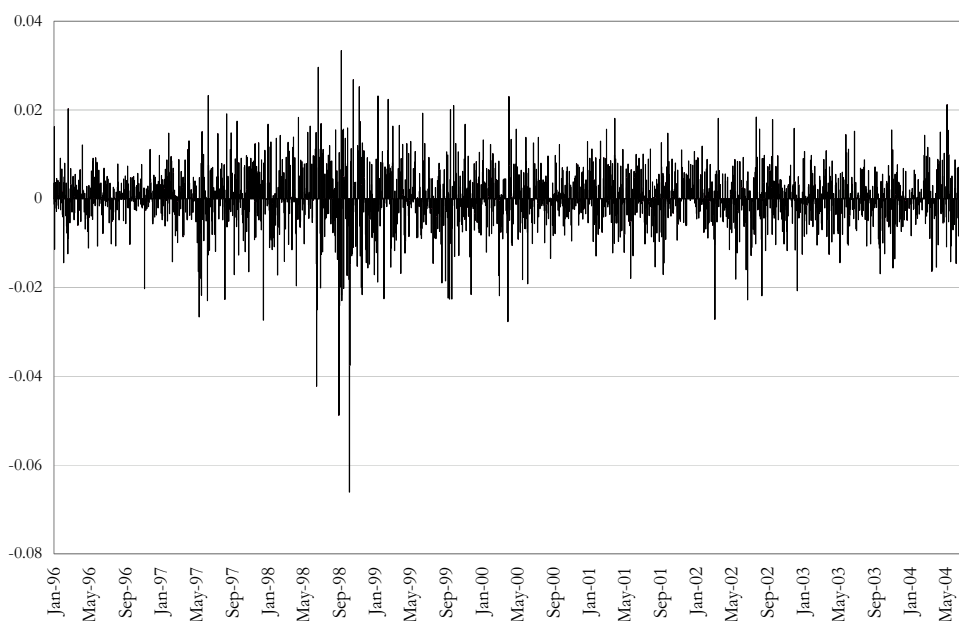
Table III.A.3.2 highlights the stylised facts or characteristics commonly observed in financial returns:

- The mean of the daily returns is very close to zero.
- Risk is expressed in two ways – standard deviation per day and volatility, which is the annualised standard deviation. If we assume returns are i.i.d. then the square root of time rule applies. This rule is explained in Section II.B.5.3. It states that the standard deviation

of b -day returns is \sqrt{b} \times standard deviation of one-day returns;⁶ or equivalently, that volatility is constant.⁷ Note that throughout this Handbook we adopt the standard convention of quoting volatility on an annual basis.

- The skewness (see Section II.B.5.5) is estimated using the Excel function ‘SKEW’ and is found to be negative.
- The excess kurtosis (see Section II.B.5.6) is estimated using the Excel function ‘KURT’ and is found to be positive, indicating that the distribution has heavy tails relative to the normal.

Figure III.A.3.1: USD/JPY returns, Jan. 1996 to July 2004



The implication of negative skewness and positive excess kurtosis is that the true probability of a large negative return is greater than that predicted under the normal distribution. This finding has potentially grave consequences for the measurement of VaR at high confidence levels. If VaR is calculated under the assumption of normality, yet the actual data have heavy tails, then VaR will understate the true risk of a disastrous outcome. Consequently, capital reserves may be an insufficient buffer to withstand disaster at the desired confidence level.

⁶ To be precise, the rule is based on *log* returns, defined as $\ln(P_{t+b}/P_t)$ where P_t is the price at time t . Log returns have the nice property that the sum of b consecutive one-day log returns is the b -day log return. Also, if b is small, it is easy to show that log returns are very close indeed to the ‘absolute’ return $(P_{t+b} - P_t)/P_t$.

⁷ To see this, let σ denote the standard deviation of one-day returns. Then the volatility based on 1-day returns = $\sigma\sqrt{250}$. Under the square root of time rule, the standard deviation of b -day returns is $\sigma\sqrt{b}$, and since there are $250/b$ time periods of length b days per annum, the volatility based on b -day returns = $\sigma\sqrt{b} \times \sqrt{(250/b)}$ which also equals $\sigma\sqrt{250}$. So in Table III.A.3.2, volatility = $0.00718 \times \sqrt{250} = 0.1135$.

Table III.A.3.2 Analysis of USD/JPY returns, Jan. 1996 to July 2004

Number of observations	2161
Mean return	2.60047×10^{-5}
Standard deviation per day	0.718%
Volatility	11.35%
Skewness	-0.8255
Excess kurtosis	6.241
No. of observations below the lower bound of 99% confidence interval	48 vs. 22 expected

Under the normal distribution, the lower bound of the 99% one-tailed confidence interval is defined by the mean less 2.33 standard deviations. In this case the lower bound equals $-2.33 \times 0.00718 = -0.0167$, so we should expect that only 1% of the return observations will be lower than this figure. With 2161 observations, only 22 returns (1% of 2161) should lie in this lower tail. In fact, examination of these data reveals that 48 returns are below the lower bound.

This kind of analysis is often performed to demonstrate that finance data violate the assumption of normality. The problem with this analysis, however, is that it incorrectly assumes that volatility has remained constant for the entire sample period of 8.5 years! Reviewing Figure III.A.3.1, we can see that the volatility of USD/JPY returns has varied considerably, with the period of greatest volatility being 1998. This was a time when most financial markets exhibited high risk following the Russian crisis. In 1998 we see that oscillations about the mean were much greater than at other times. High-volatility periods are characterised by large returns, both positive and negative. Note that it is the absolute size of the return that is important rather than direction. On one day at the height of the Russian crisis, the US dollar depreciated by more than 6% against the Japanese yen. The largest positive return (in excess of 3%) occurred during the same period of high volatility.

Of the 48 observations in the lower tail of the confidence interval, 18 fall in 1998, a period of very high volatility. This leads us to an alternative way of understanding the data: the large number of observations that appear in the lower tail is a result of changes in volatility throughout the sample period. Sometimes, as in 1998, volatility is much higher and so the confidence interval is correspondingly wider. Instead of a lower bound of -0.0167 , the lower bound would be very much lower (for example, -0.0335 if volatility were to double). If we take account of the increase in volatility it is likely we will find that the number of observations in the tail is close to 1% as expected. In other words, the main problem with financial data is not skewness/kurtosis (often referred to as heavy tails), but the fact that volatility is changing over time. In short, if we

take account of these changes in volatility then the assumption of normality may actually be quite a reasonable one.

This brings us to a further discussion of ‘i.i.d.’ –returns.⁸ If returns are ‘identically distributed’, then the parameters of the distribution (mean and variance) should be constant over time. This assumption is clearly violated since the variance parameter is changing. If returns are independent, then yesterday’s return will have no bearing on today’s return. Tossing coins is a good example of independent outcomes. Even if I toss 10 heads in a row, the probability of heads on the next toss is still 50% (assuming a fair coin!). In other words, the next toss is unaffected by what has gone before.

However, the empirical analysis of financial returns shows that the assumption of independence is unsupported: the *size* (but not the direction) of yesterday’s return *does* have implications for today’s return. A large return yesterday (in either direction) is likely to be followed by another large return in either direction. This concept is commonly referred to as ‘the heat wave effect’ or ‘volatility clustering’. We can test for volatility clustering by examining the autocorrelation in squared returns. Squared returns are used because squaring removes the sign of the return – we can focus purely on its magnitude rather than its direction. Financial data often generally exhibit significant positive autocorrelation in squared returns, in which case the data are not i.i.d.

Volatility clustering is arguably the most important empirical characteristic of financial data. The most useful financial models take account of this fact. Indeed, the world of finance research was forever changed in the 1980s when volatility clustering was first identified by Robert Engle and his associates. In 2003 Robert Engle won (jointly) the Nobel Prize for Economics, largely because of his groundbreaking contribution in this area.

To summarise this section, we can say that traditional financial models have often assumed that returns are i.i.d. normal. In reality we observe changes in volatility and volatility clustering. We will show in the following sections that these characteristics can be modelled successfully. Having taken account of volatility clustering, the issue of heavy tails becomes less significant.

⁸ See Section II.B.5.3 for further discussion of this concept.

III.A.3.3 Models of Volatility Clustering

If we believe that today's volatility is positively correlated with yesterday's volatility, then it is appropriate to estimate *conditional* volatility, that is, volatility that is conditional on the recent past. We will discuss two methods for achieving this: exponentially weighted moving average (EWMA) and generalised autoregressive conditional heteroscedasticity (GARCH). The latter is more difficult to implement but offers some potential advantages. The application of these methods to VaR models will be discussed in Section III.A.3.4.

III.A.3.3.1 Exponentially Weighted Moving Average

The EWMA method for estimating volatility was popularised by the RiskMetrics Group. Today's estimate of variance is:

$$\hat{\sigma}_t^2 = (1 - \lambda)r_{t-1}^2 + \lambda\hat{\sigma}_{t-1}^2, \quad (\text{III.A.3.1})$$

where λ is a 'smoothing constant' and r_{t-1} is the most recent day's return.⁹ Notice that since λ is positive, today's variance will be positively correlated with yesterday's variance, so we see that EWMA captures the idea of volatility clustering. The parameter, λ , may also be referred to as the 'persistence' parameter. The higher the value of λ , the more will high variance tend to persist after a market shock. The EWMA variance also reacts immediately to market shocks. If yesterday's return is large, in either direction, the variance will increase through the first term on the right-hand side of (III.A.3.1). The greater is $1 - \lambda$, the greater will be the size of the reaction to a return shock.

In Section III.A.3.3.2 we shall see that this type of reaction to, and persistence following, a market shock is the characteristic of a GARCH model. In fact, the EWMA can be thought of as a very simple GARCH model. However, whilst GARCH volatility models are based on firm statistical foundations, EWMA volatility models are not, for the following reasons:

- There is no proper statistical estimation procedure for the smoothing constant: the user simply assumes some value for λ . According to RiskMetrics, a value for λ of around 0.94 is generally appropriate when analysing daily data.¹⁰

⁹ This can be calculated in Excel by typing (III.A.3.1) in the formula bar, or by using the exponential smoothing analysis tool. Note that technically, in (III.A.3.1) the volatility estimate depends on the entire historical data set rather than a limited past history. A pragmatic alternative calculation that is much easier to replicate for the auditors is to truncate the historical sample at n past observations, where n is defined as the point at which λ^n drops below some critical value C . For instance, if $\lambda = 0.94$ and $C = 0.002$ then $n = 100$ observations. The EWMA on n observations can be written as a finite sum:

$$\hat{\sigma}_t^2 = \left[\sum_{i=1}^n (1-\lambda)^i r_{t-i}^2 \right] / \left[\sum_{i=1}^n (1-\lambda)^i \right]$$

¹⁰ See Section 5.3.2 of RiskMetrics Technical Document, 1996, available at: www.riskmetrics.com/techdoc.html#rmttd

- In portfolio models, where the variance is calculated using a covariance matrix, some portfolios can have negative variance unless this matrix is positive semi-definite (see Section II.D.3.2). For this reason one cannot form an EWMA covariance matrix using different values for λ for different assets: in fact the *same* value of λ must be used for all variances and covariance in the matrix. More details are given in Section III.A.3.4.1.
- The EWMA variance estimate is converted to volatility by taking the square root, to obtain the standard deviation, and then applying the square root of time rule:

$$EWMA \text{ volatility estimate at time } t \text{ for a horizon of } b \text{ days} = \hat{\sigma}_t \sqrt{b} \times \sqrt{250/b}.$$

But, as explained in Section III.A.3.2, this is equivalent to a *constant* volatility assumption. Hence the EWMA model is not really appropriate for estimating the market evolution over time horizons longer than a few days.

Figure III.A.3.2: EWMA volatility – USD/JPY returns, Jan. 1996 to July 2004

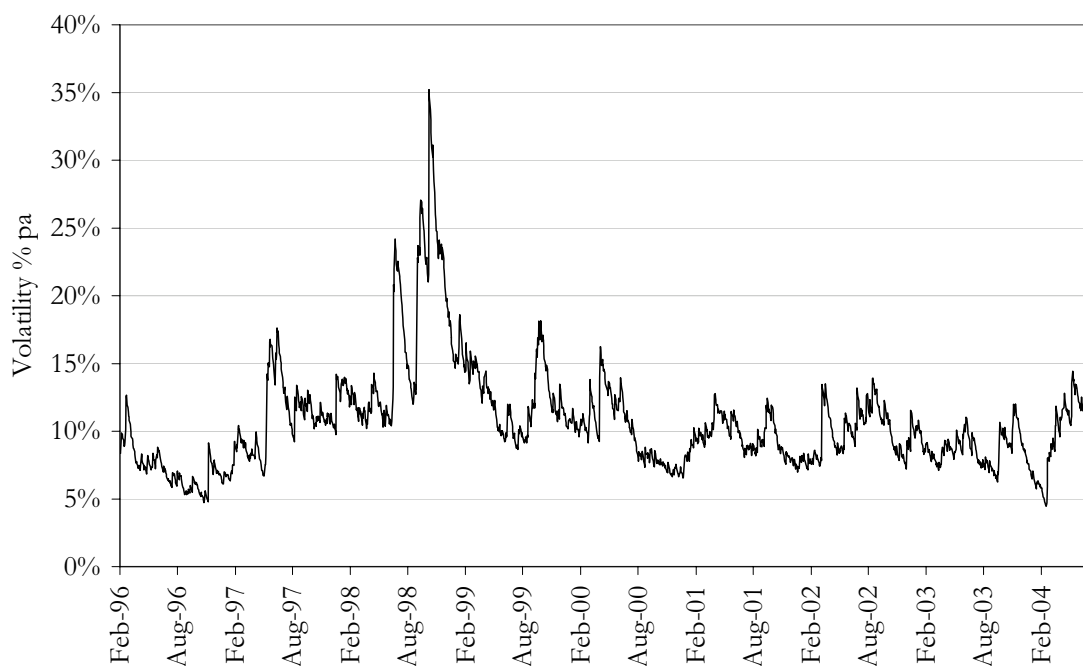


Figure III.A.3.2 highlights the variability of conditional volatility. Here the EWMA model has been applied with $\lambda = 0.94$. Recall from the previous section that unconditional volatility for this same time period was calculated as 11.35% pa. Following the market shocks of 1998, conditional volatility reaches a maximum of 35%. In the quieter periods of 1996 and early 2004, conditional volatility briefly dips below 5% pa.

Having estimated conditional volatility in this way, we can then standardise the daily returns. That is, each daily return is divided by the relevant standard deviation. We do this to try to make

the returns comparable. Prior to standardisation, each return comes from a distribution having a different volatility and is therefore not strictly comparable. We then repeat the analysis of Table III.A.3.2, but this time using the standardised returns (see Table III.A.3.3).

Table III.A.3.3 Analysis of standardised USD/JPY returns, (Feb. 1996 to July 2004)

Number of observations	2140
Skewness	-0.1974
Excess kurtosis	0.2727
No. of observations below the lower bound of 99% confidence interval	29 vs. 21 expected

Having standardised returns on conditional volatility, we find that the problems noted earlier are much diminished. That is, skewness and excess kurtosis are now much closer to zero. The number of observations in the lower tail is much closer to that expected under normality. This suggests that the issue of heavy tails, often noted by finance analysts, can be partly explained by volatility clustering.

III.A.3.3.2 Generalised Autoregressive Conditional Heteroscedasticity Models

GARCH models are similar to EWMA in that both focus on the issue of volatility clustering. The word heteroscedasticity is Greek for ‘different scale’, so GARCH models are concerned with the process by which the scale of returns, or volatility, is changing. GARCH models are ‘generalised’ in the sense that they can be varied, almost infinitely, to take account of the factors specific to a particular market. What all GARCH models share, however, is a positive correlation between risk yesterday and risk today; that is, an ‘autoregressive’ structure in risk.

The simplest GARCH model consists of two equations that are estimated together. The first is the conditional mean equation and the second is for the conditional variance:

$$\begin{aligned}
 r_t &= \mu + \varepsilon_t \\
 \sigma_t^2 &= \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \\
 \omega > 0, \alpha, \beta &\geq 0.
 \end{aligned}
 \tag{III.A.3.2}$$

The parameters ω , α and β of the GARCH model are usually estimated using maximum likelihood estimation (MLE). This involves using an iterative method to find the maximum values of the likelihood function. In normal GARCH models, i.e. when the distribution of ε_t conditional on all information up to time t is normal with variance σ_t^2 , the likelihood function is a multivariate normal distribution on the model parameters (see Section II.E.4.7).

In its simplest form (shown here) the conditional mean equation merely adjusts for the mean (c), leaving an ‘unexpected’ return ε_t . The conditional variance equation appears very similar to the EWMA equation where β stands in place of λ , α stands in place of $1 - \lambda$, and an extra constant term (ω) is also included. Perhaps the most important difference between EWMA and GARCH is the fact that in GARCH there is no constraint that the sum of the coefficients ($\alpha + \beta$) should equal one. It is possible that data are best explained when the estimated coefficients do add to one,¹¹ but in this case the GARCH forecasts behave just like those from a constant volatility model. If the sum $\alpha + \beta$ is less than one (the more usual case) then volatility is said to be mean-reverting and the rate of mean reversion is inversely related to this sum. This means that the variance will, in the absence of a market shock, tend towards its steady-state variance defined by

$$\sigma^2 = \frac{\omega}{1 - \alpha - \beta}. \quad (\text{III.A.3.3})$$

The GARCH volatility forecasts then behave like the volatility term structures we observe in implied volatilities, where implied volatilities of long-term options do not vary as much as the implied volatilities of short-term options. From (III.A.3.3) we can see that if $\alpha + \beta = 1$ (as in EWMA) then the denominator equals zero so the steady-state variance is undefined. In this case variance is not mean-reverting and is assumed constant as we project forwards in time.

Example III.A.3.1: GARCH model for spot USD/JPY

We estimate a GARCH(1,1) model for daily log returns from January 1996 to July 2004. The conditional variance equation is estimated using a maximum likelihood technique:

$$\sigma_t^2 = 3.78 \times 10^{-7} + 0.03684\varepsilon_{t-1}^2 + 0.95571\sigma_{t-1}^2.$$

The constant term (3.78×10^{-7}) is statistically significant but very small. Conditional variance for the USD/JPY is quite persistent (with a persistence coefficient of 0.955710) and not particularly reactive (reaction coefficient of 0.03684) compared with some other markets. This means that the initial reaction to new information will be more muted, but its effect relatively long-lasting. As the sum of these two coefficients is less than unity, we can say that conditional variance is mean-reverting to a steady-state variance defined by (III.A.3.3). Substituting the estimated parameters into this equation, the steady-state variance is 0.0000505794, equivalent to an annualised volatility of 11.24%. This is close to the sample volatility of 11.35% p.a. in Table III.A.3.2.

¹¹ When this happens, the model is called an integrated GARCH (IGARCH) model. The IGARCH is most commonly used for modelling currency returns.

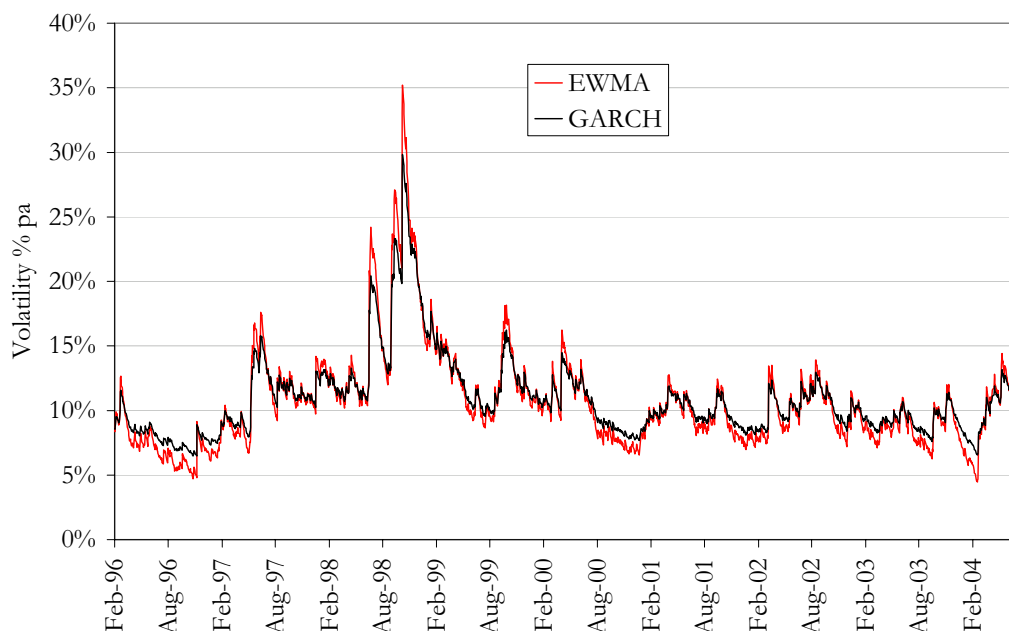
Figure III.A.3.3: GARCH vs. EWMA volatility – USD/JPY returns

Figure III.A.3.3 shows the conditional volatility resulting from the GARCH model. The two series, EWMA and GARCH, track each other closely, so the figure emphasises the similarity between the two approaches.

We have argued that volatility clustering is the main cause of the apparent heavy tails observed in financial data. If GARCH modelling is successful, then it should account for these heavy tails, leaving residuals that are i.i.d. normal. Is this in fact the case? It can be tested by examining the series of *standardised* returns. To obtain the standardised returns we estimate a GARCH model and from this create a series of daily conditional standard deviation estimates. We then divide each daily return by the relevant conditional standard deviation to obtain a standardised return, which we then test for normality. If this series is normally distributed then we can conclude that volatility clustering fully explains the extreme moves. If not, then further action may be necessary to adjust for the extreme moves (or for the heavy tails of the distribution). For example, it may be necessary to use a more complex GARCH specification. For instance, we might use an asymmetric specification for the conditional variance that can account for a ‘leverage effect’ in equities, and/or a conditional distribution for ϵ_t that is non-normal. There is a huge body of research literature on GARCH models, but this is beyond the scope of the PRM exam.

III.A.3.4 Volatility Clustering and VaR

Volatility clustering has important implications for VaR. Consider the situation in which new and unexpected information comes to the market, as it did when Russia defaulted on its debt in 1998, causing a large reaction in market prices. Our knowledge of volatility clustering tells us that this market shock is likely to be followed by large returns (in either direction) for some time. Ideally, our VaR measure will increase significantly, sending the appropriate signal to risk managers either to reduce risk through hedging or to ensure that capital is adequate to withstand the higher risk environment.

Figure III.A.3.4: USD/JPY volatility in 1998

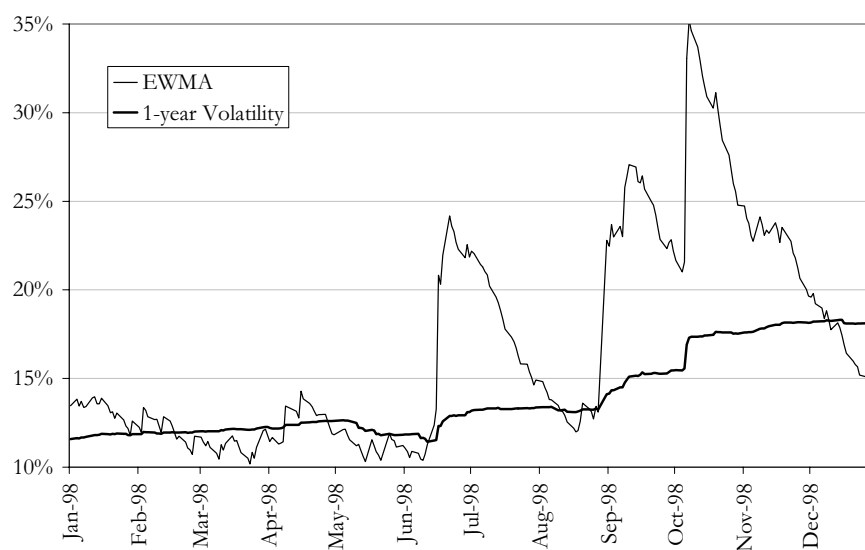


Figure III.A.3.4 compares two volatility measures for USD/JPY returns for calendar year 1998. The EWMA measure of conditional volatility reacts swiftly as the crisis unfolds in mid-1998. The other series shows unconditional volatility calculated using rolling one-year samples. Since each day in the sample is equally weighted, and has a small weight of only $1/250$, this measure is very slow to react to market shocks.

The choice of volatility measure will have major implications for the VaR measure and consequently for capital adequacy. By failing properly to take account of volatility clustering, the risk is that financial institutions will take unduly large risks (or will hold insufficient capital) in periods of market crisis. In addition, they will hold too much expensive capital at other times.

Backtests of such models will reveal clusters of exceptions where the actual losses exceed VaR. Unfortunately, the timing of exceptions is often ignored in traditional backtesting procedures which focus only on the *number* of exceptions (see Section III.A.2.8). Both regulators and risk

managers are concerned about the timing and size of exceptions. A series of large losses in quick succession is potentially far more serious for solvency than smaller losses spread over time, even if the total combined losses are equal.

For reasons that remain unclear to the authors, the Basel regulations relating to market risk currently require financial institutions to measure volatility using at least one year of data. This regulation encourages the use of volatility measures that react very slowly to new information such as the one illustrated above. In our view, this is exactly the wrong way to proceed.¹²

III.A.3.4.1 VaR using EWMA

Volatility clustering can be relatively easily incorporated into VaR measures using the EWMA approach. There are at least three ways that this can be done:

- *Historical simulation using volatility weighted data.* This method is explained in detail in Section III.A.2.6.2. Historical returns are standardised using conditional volatility estimates calculated using EWMA. This approach has a number of attractions, especially for option-affected portfolios. Historical simulation makes no assumption about the distribution of historical returns apart from independence. Thus it is attractive if it is feared that the standardisation process has not entirely eliminated the heavy tails evident in raw returns.
- *MC simulation using EWMA.* Returns could be simulated under the assumption of normality, but using a covariance matrix created using EWMA. Again, this method is most relevant for option-affected portfolios.
- *Analytical VaR using EWMA.* We will explain this method here, building on the discussion in Section III.A.2.4.

Of these three methods, the last two make use of the assumption that returns are conditionally normally distributed. In Section III.A.3.3 we explained that normality is a much more reasonable assumption to make if we properly account for changes in volatility. One way to do this is to calculate the covariance matrix using the EWMA measures of variance and covariance. This is preferable to the standard measures of unconditional variance and covariance, which are slow to respond to new information when estimated using long samples.¹³

Equation (III.A.3.1) shows the EWMA equation for variance. The analogous equation for covariance between assets 1 and 2 is:

¹² See the discussion of 'historical observation period' in Basel Committee 'Amendment to the Capital Accord to Incorporate Market Risks', January 1996, at p. 44.

¹³ See Sections II.B.5 and II.B.6 for descriptions of these standard measures.

$$\hat{\sigma}_{12,t} = (1 - \lambda)r_{1,t-1}r_{2,t-1} + \lambda\hat{\sigma}_{12,t-1},$$

where $r_{1,t-1}$ and $r_{2,t-1}$ are yesterday's returns for assets 1 and 2, respectively. Note that when constructing a large covariance matrix it is always important to ensure that it is positive semi-definite, otherwise portfolio volatility may not be defined (see Section II.D.3.2). If we use a different value of λ for each variance and covariance term, the matrix will not necessarily be positive semi-definite. RiskMetrics gets around this problem by using the same value for λ (being 0.94 in the case of daily data) throughout the covariance matrix.

When calculating VaR we need to *forecast* portfolio variance for the horizon of interest, say, 10 business days. Unlike GARCH, EWMA is a non-stationary model of variance. That is, under EWMA, variance is *not* mean-reverting. This means that the EWMA forecast of any future variance (or covariance) is the same as the estimate made today. Equivalently, as explained in Section III.A.3.3.1, the average volatility over any forecast horizon is a constant, equal to the volatility estimated today.

Once the covariance matrix has been defined, it can then be used for VaR calculations using either:

- the analytical method (appropriate for simple linear portfolios) or
- MC simulation (best for option-affected portfolios).

In the analytical method (See Section II.A.2.4) the b -day VaR estimate at the significance level α is given by:

$$\text{VaR}_{\alpha,b} = Z_{\alpha}P\sigma, \tag{III.A.3.4}$$

where Z_{α} is the standard normal α critical value, P is the current value of the portfolio and σ is the forecast of the standard deviation of the b -day portfolio return.

The standard deviation in (III.A.3.4) is computed using a forecast covariance matrix of b -day returns as follows (see Section II.D.2.1):

(i) *Representation at the asset level,*

$$\sigma = \sqrt{\mathbf{w}'\mathbf{V}\mathbf{w}}$$

where $\mathbf{w} = (w_1, \dots, w_n)$ is the portfolio weights vector and \mathbf{V} is the b -day forecast of the covariance matrix of *asset* returns.¹⁴

¹⁴ Equivalently, $\text{VaR}_{\alpha,b} = Z_{\alpha}\sigma_{P\&L} = Z_{\alpha}\sqrt{\mathbf{p}'\mathbf{V}\mathbf{p}}$, where $\mathbf{p} = P\mathbf{w}$ is the vector of nominal amounts invested in each asset.

(ii) Representation at the risk factor level,

$$\sigma = \sqrt{\boldsymbol{\beta}'\mathbf{V}\boldsymbol{\beta}}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n)$ is the portfolio sensitivity vector and \mathbf{V} is the b -day forecast of the covariance matrix of *risk factor* returns.¹⁵

To calculate *analytic* VaR using EWMA we simply calculate today's portfolio variance as above. To convert to a 10-day horizon we simply multiply the one-day VaR by $\sqrt{10}$; or we use a 10-day covariance matrix for \mathbf{V} . In either case, the results will be the same. The 10-day matrix can be obtained, using the square root of time rule, by multiplying every element of the one-day covariance matrix by 10. But note that these two conversion methods assume both constant volatility and no serial correlation in the forward projection of volatility, which is one reason why the EWMA VaR estimate does not fully reflect volatility clustering.

By contrast, for *non-linear* portfolios the MC VaR method uses a covariance matrix, and you should be sure to use the 10-day covariance matrix directly in the simulations. You will get an incorrect result if you simulate the one-day VaR and multiply the result by $\sqrt{10}$.

Example III.A.3.2: Analytical method with EWMA for two-asset portfolio

Consider a simple portfolio with \$1m invested in asset 1 and \$2m invested in asset 2. Suppose the EWMA model estimated on daily returns for asset 1 gives a variance estimate of 0.01, for asset 2 returns the EWMA variance is 0.005 and the EWMA covariance is 0.002. What is the 5% 10-day VaR?

We have $\mathbf{p} = (1, 2)'$ and \mathbf{V} is the matrix $\begin{pmatrix} 0.01 & 0.002 \\ 0.002 & 0.005 \end{pmatrix}$. So the P&L volatility = $\sqrt{\mathbf{p}'\mathbf{V}\mathbf{p}} = \sqrt{0.038} = \0.19m and the 5% one-day VaR is $1.645 \times 0.195 = \$0.3207\text{m}$. Thus the 5% 10-day VaR = $0.3207 \times \sqrt{10} = \1.014m .¹⁶

This is high, because the assets have a very high variance (and covariance). For instance, a daily variance of 0.01 corresponds to an annual volatility of $\sqrt{2.5} = 158\%$! In fact, the 1% 10-day VaR for this portfolio is $2.33 \times 0.195 \times \sqrt{10} = \1.43m , so almost half the amount invested would be required for risk capital to cover this position! However, the example illustrates an important

¹⁵ Equivalently $\text{VaR}_{\alpha,b} = Z_{\alpha}\sigma_{\text{P\&L}} = Z_{\alpha}\sqrt{\mathbf{p}'\mathbf{V}\mathbf{p}}$, where $\mathbf{p} = P\boldsymbol{\beta}$ is the vector of sensitivities to each risk factor in nominal terms.

¹⁶ Of course, the same result would be obtained if \mathbf{V} is the given matrix but with each element multiplied by 10 and we calculated $Z_{\alpha}\sqrt{\mathbf{p}'\mathbf{V}\mathbf{p}}$.

weakness in the use of the analytic VaR method for long-only positions. Being based on the assumption that portfolio P&L is normally distributed, there is a chance, however small, that the estimated VaR will be *more* than the total investment. For a long-only position, this clearly is a nonsensical result.

Example III.A.3.3: Analytical method for portfolio mapped to two risk factors

Suppose a US investor buys \$2m of shares in a portfolio of UK (FTSE100) shares, and the portfolio beta is 1.5. Suppose the FTSE100 and GBP/USD volatilities are 15% and 20% respectively (with corresponding variances of $0.15^2 = 0.0225$ and $0.20^2 = 0.04$) and their correlation is 0.3. What is the 1% 10-day risk factor VaR in US dollars?

We have two risk factors (FTSE and GBP/USD) with $\mathbf{p} = (3, 2)'$. Note that the \$2m exposure to the equity portfolio described above is equivalent in risk terms to a \$3m exposure to the FTSE index since beta is 1.5. The one-day variances are: $0.0225/250 = 0.00009$ for the FTSE and $0.04/250 = 0.00016$ for GBP/USD. With a correlation of 0.3, the one-day covariance is $(0.3 \times 0.15 \times 0.2)/250 = 0.000036$. Hence,¹⁷

$$\mathbf{p}'\mathbf{V}\mathbf{p} = \begin{pmatrix} 3 & 2 \end{pmatrix} \begin{pmatrix} 90 & 36 \\ 36 & 160 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} \times 10^{-5} = \begin{pmatrix} 342 & 428 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} \times 10^{-5} = 0.01882$$

and the 1% 10-day VaR is therefore $2.32634 \times \sqrt{10} \times 0.01882 = \0.31845m .

In concluding this section, we can say that VaR calculations, whether analytical or simulation-based, can be greatly improved by taking account of volatility clustering. In this regard EWMA is far superior to the approach, unfortunately encouraged by regulators, which employs unconditional volatility based on large samples of data.

III.A.3.4.2 VaR and GARCH

Alternatively, the issue of volatility clustering may be incorporated into VaR estimates using the more sophisticated GARCH approach. While GARCH is undoubtedly more challenging to implement than some other methods, it also has some distinct advantages. For example, because GARCH is a more general model it can explain the characteristics of the data more precisely to ensure that all evidence of non-normality and dependence is removed from the standardised returns.

¹⁷ Note that the one-day covariance matrix is $\begin{pmatrix} 0.00009 & 0.000036 \\ 0.000036 & 0.00016 \end{pmatrix}$ so the 10-day covariance matrix is

$$\begin{pmatrix} 0.0009 & 0.00036 \\ 0.00036 & 0.0016 \end{pmatrix} = \begin{pmatrix} 90 & 36 \\ 36 & 160 \end{pmatrix} \times 10^{-5}.$$

A distinct advantage of GARCH over EWMA is that GARCH is a mean-reverting model of volatility. This fits better with the stylised facts observed in the market; we regularly observe that volatility will tend toward a ‘mean’ or ‘steady-state’ value after a period of unusually high or low risk. If we are forecasting volatility over, say, the next 10 days, then it could be helpful to take account of this mean reversion feature. We cannot do this with EWMA volatility models because they use the square root of time rule, as explained above.

Example III.A.3.4: Forecasting volatility with GARCH

Suppose that we have estimated a GARCH model for a stock index such that the mean return is zero and the conditional variance equation is as follows:

$$\sigma_t^2 = 5.0 \times 10^{-6} + 0.07\varepsilon_{t-1}^2 + 0.88\sigma_{t-1}^2$$

In this case the steady-state variance (using (III.A.3.3)) is equal to 0.0001, equivalent to daily volatility of 1% and annual volatility of $1\% \times \sqrt{250} = 15.81\%$ p.a. Suppose that we are estimating VaR at a time when the market has recently been unusually quiet. The current estimate of unconditional annual volatility (using, say, one year of daily data) stands at 13.0% and today’s estimate of conditional daily volatility is 0.007746, equivalent to annual volatility of $0.7746\% \times \sqrt{250} = 12.25\%$. Assume that today new and unexpected economic data hit the market, causing a large return of +0.04 or 4%. To put this in perspective, a return of 4% in one day is a greater than 5 standard deviation event (using unconditional volatility).

To forecast variance tomorrow we proceed as follows, substituting the appropriate values for today’s shock (0.04) and today’s variance (0.00006):

$$\sigma_{t+1}^2 = 5.0 \times 10^{-6} + (0.07 \times 0.04^2) + (0.88 \times 0.00006) = 0.0001698$$

Notice the large size of forecast variance (0.0001698 vs. 0.00006) as variance reacts to today’s market shock. When forecasting variance on the subsequent day (and thereafter) we do not know the return shock so variance is forecast as:

$$\sigma_{t+2}^2 = 5.0 \times 10^{-6} + (0.07 + 0.88) \times 0.0001698 = 0.0001663$$

and similarly for the days afterwards. Table III.A.3.4 shows the daily variance forecasts for each day in the 10-day horizon.

While tomorrow’s forecast volatility is significantly higher than today’s, thereafter the forecast volatility falls slightly. The GARCH model tells us that in the absence of any further shock, volatility will gradually revert towards the steady-state volatility. We obtain a forecast of volatility

over the next 10 days by summing the 10 daily variances, multiplying by 250/10 and taking the square root. This gives 19.7% pa. In contrast, the unconditional volatility forecast for the next 10 days is approximately 13% – today’s large return having only a marginal impact due to its small weighting of 1/250.

Table III.A.3.4 Forecasting Volatility with GARCH

Day	Variance	Equivalent volatility % pa
T+1	0.0001698	20.6%
T+2	0.0001663	20.4%
T+3	0.0001630	20.2%
T+4	0.0001598	20.0%
T+5	0.0001569	19.8%
T+6	0.0001540	19.6%
T+7	0.0001513	19.4%
T+8	0.0001487	19.3%
T+9	0.0001463	19.1%
T+10	0.0001440	19.0%
10-day horizon i.e. T+1 to T+10	Sum = 0.0015601	19.7%

The significance of this for the VaR estimate is obvious; using the GARCH conditional volatility forecast will significantly boost the VaR, signalling to risk managers that risk should either be substantially reduced or capital increased to withstand the new high-risk environment.

This example also illustrates the point that the square root of time rule is inappropriate in a world of volatility clustering. The GARCH model tells us that in this situation, volatility is likely to increase initially and then gradually decline over the 10-day horizon. To assume constant volatility for 10 days is not suitable.

For the professional risk manager calculating VaR, the task of implementing GARCH techniques might appear daunting. How might she actually go about it? There are a number of possibilities:

- *Analytical VaR.* As explained in Section III.A.3.4.1, but this time using a covariance matrix based on GARCH variance–covariance forecasts over the risk horizon. This is a convenient solution for linear portfolios.
- *Historical simulation using volatility weighted data.* As explained in Section III.A.3.4.1, but this time standardising returns using GARCH volatility estimates. This could be done quite simply, using a univariate GARCH model for each asset/factor. This method makes no assumption about the distribution of standardised returns apart from independence.
- *Monte Carlo simulation.* A GARCH covariance matrix forecast could be used to simulate returns going forward. This would allow for changes in volatility, as well as the

underlying, over the risk horizon – an important advantage for portfolios containing options.

Professional risk managers are generally analysing investment or trading portfolios containing multiple assets. To use some of the GARCH methods listed above it is necessary to evaluate an entire covariance matrix. This can present substantial implementation problems. The difficulty compounds as the number of assets or risk factors grows, particularly as we must ensure that the covariance matrix is positive semi-definite. Various approaches have been suggested in the academic literature, and these are surveyed by Bauwens *et al.* (2003). One very simple alternative is to use GARCH volatilities but assume a constant correlation between assets or risk factors, as proposed by Bollerslev (1990). Variance terms in the matrix are estimated using the simple GARCH(1,1) specification as shown in (III.A.3.2). The covariance terms in the matrix are formulated as follows:

$$\sigma_{ij,t+1} = \rho_{ij} \sigma_{i,t+1} \sigma_{j,t+1}$$

where ρ_{ij} is the sample correlation of mean-corrected returns. Ease of estimation is achieved by constraining the correlation to be constant over the estimation period and by ignoring cross-market effects in volatility.

The benefits of the GARCH approach to VaR estimation have recently been illustrated by Berkowitz and O'Brien (2002). They have shown that a simple reduced-form GARCH implementation produces regulatory capital estimates that are an improvement on the methods used by some large commercial banks. The reduced-form GARCH approach applies a univariate GARCH model directly to portfolio P&L data, thus avoiding the need for a large covariance matrix. As in the standard historical simulation approach described in Chapter III.A.2, an artificial history of daily P&Ls is created using the current portfolio weights. But instead of taking a lower percentile of the empirical distribution of these P&Ls for the VaR estimate, they apply a GARCH model to the portfolio returns series and use formula (III.A.3.4). They find that the VaR estimates based on this type of GARCH model are more sensitive to changes in volatility; they allow for more risk-taking (or less capital) when volatility is low and less risk-taking (or more capital) when volatility is high. These excellent results are achieved without adversely affecting the number of exceptions in backtests (see Section III.A.2.8 for a description of VaR model backtests); in fact the size of the maximum exception is reduced.

In summary, GARCH techniques have the potential to greatly enhance our modelling of market risk and to ensure that appropriate capital buffers are in place.

III.A.3.5 Alternative Solutions to Non-normality

Here we consider some approaches to estimating VaR in the face of non-normality that differ from those considered in Section III.A.3.4 in that they do not take account of volatility clustering. We limit ourselves to discussing only three possibilities, although many others exist.

III.A.3.5.1 VaR with the Student's t distribution

The Student's t distribution is often proposed as a possible candidate for describing financial returns because of its heavy tails. It is actually a poor candidate because it assumes returns are i.i.d. Under the Student's t distribution each day's return is assumed independent of the previous day's return; we know this is unlikely to be the case because of the heat wave effect. The implication of this is that the t distribution will tend to underestimate VaR in periods of market crisis – the time when risk measurement is most crucial – and overestimate VaR when market conditions are quiet. Backtests will reveal clusters of exceptions where actual losses exceed VaR.

Nevertheless the Student's t distribution remains quite popular with some professional risk managers. Statistical background is provided in Section II.E.4.6. The standard Student's t has only one parameter, ν , the 'degrees of freedom'. The distribution was originally designed for working with small samples where the degrees of freedom are one less than the sample size. As ν approaches infinity, the distribution converges to the normal.

Under the standard Student's t distribution:

- the mean is equal to zero,
- the variance is equal to $\frac{\nu}{\nu - 2}$,
- the skewness is equal to zero and
- the (raw) kurtosis is equal to $\frac{3(\nu - 2)}{\nu - 4}$.

In VaR applications we will be working with large data sets and attempting artificially to select the parameter ν to fit the shape of the tails of the distribution (that is, to match the sample kurtosis).¹⁸ Since the observed variance will not be equal to $\nu/(\nu - 2)$ it will be necessary to scale the variance.¹⁹ It will generally not be necessary to scale the mean as mean returns (at daily or higher frequencies) are close to zero. Dowd (2002) explains how to adapt the standard Student's t distribution for VaR calculations (for a single asset) as follows:

¹⁸ It is also possible to estimate the degrees of freedom parameter using maximum likelihood techniques.

¹⁹ This is analogous to the way in which we adapt the standard normal distribution. The standard normal has variance and standard deviation of one. When we use data having an observed standard deviation different from one, we divide the observed standard deviation by unity.

- (a) Select the degrees of freedom parameter by matching it to the sample kurtosis such that:

$$(Raw)Kurtosis = \frac{3(\nu - 2)}{(\nu - 4)}. \quad (III.A.3.5)$$

- (b) The empirical variance should be scaled by:

$$\frac{\nu - 2}{\nu}. \quad (III.A.3.6)$$

- (c) Select the appropriate critical point from the t distribution, based on the desired level of probability (e.g. 0.01) and the degrees of freedom selected in (a).
- (d) Proceed with VaR calculation using, for instance, the analytical method – but the MC VaR model could also be applied.

Example III.A.3.5: VaR with Student's t

For purposes of comparison we use the USD/JPY data described in Section III.A.3.2. The returns have empirical *excess* kurtosis of 6.241, equivalent to *raw* kurtosis of 9.241. Using equation (III.A.3.3) and solving for ν gives a value of 4.961. The empirical variance is scaled using equation (III.A.3.4) to give adjusted daily variance of:

$$0.00718^2 \times \frac{4.961 - 2}{4.961} = 3.07693 \times 10^{-5},$$

being equivalent to daily volatility of 0.005547. The critical value of the t distribution can be found in Excel using the TINV function.²⁰ At the 99% level of confidence and with 5 degrees of freedom, the critical value of the t distribution is 3.36 (the comparable critical value of the normal distribution is 2.33).

We use this information to calculate VaR using the analytical method for a position held long \$1m. We choose a 10-day holding period and ignore expected returns. We compare VaR for the Student's t with VaR from the more familiar normal distribution. Note that the square root of time rule is appropriate here as returns are assumed i.i.d.:

- Student's t VaR = $3.36 \times \sqrt{10} \times 0.005547 \times \$1,000,000 = \$58,938$
- Normal VaR = $2.33 \times \sqrt{10} \times 0.00718 \times \$1,000,000 = \$52,903$

²⁰ Note that the Excel TINV function assumes a two-tailed probability distribution, whereas VaR calculations generally apply a one-tailed probability distribution. When using TINV, take care to double the probability parameter to get the appropriate value for a one-tailed distribution. If you are interested in probability of 0.01 (i.e. one-tailed 99% confidence), then you should instead use a probability of 0.02. Note also that the Excel TINV function assumes that ν is an integer. In this example we use '=TINV(0.02,5)'. If we instead used '=TINV(0.02,4.961)' Excel would round 4.961 *down* to 4, returning a value of 3.75.

The Student's t VaR estimate is 11% higher than that calculated using the normal distribution. This example illustrates how the heavier-tailed Student's t distribution will tend to give larger VaR estimates than the normal. Applying the Student's t approach will tend to result in greater capital requirements over time (or reductions in the amount of risk taken).

Potentially these larger VaR estimates afford financial institutions greater protection from extreme variations as they will respond either by reducing risk or increasing capital. Given the existence of volatility clustering, however, even an 11% increase in VaR will not be sufficient protection in times of market crisis. Returning to Figure III.A.3.4, we see that in 1998, conditional volatility more than tripled, reaching a peak of 35% p.a.!

III.A.3.5.2 VaR with Extreme Value Theory

Some VaR applications focus on the 'tail behaviour' of financial returns distributions. For instance, economic capital is often estimated at an extremely high percentile, such as 99.97% for a AA company (see Section III.0.2.3). For another example, when conducting stress tests the possible behaviour of portfolio returns is pushed to an extreme limit (see Chapter III.A.4).

Several large banks have been developing risk capital models based on extreme value theory (EVT). In these models, it is not VaR but an associated risk metric called *conditional VaR* – also called 'expected shortfall', 'expected tail loss' or 'tail VaR' by various authors – that is estimated. Whereas VaR is the cut-off point, above which the largest losses occur, the associated conditional VaR is the average of these largest losses. For instance, the 1% VaR based on 1000 P&Ls is the 10th largest loss; the 1% conditional VaR is the average of these 10 largest losses.

Though not admissible under the Basel rules for the computation of regulatory capital, conditional VaR measures are commonly favoured for the computation of economic capital. That is because conditional VaR is *sub-additive*, i.e. the sum of the component conditional VaRs can never be less than the total conditional VaR (see Section III.A.3.6). Of course, any risk metric that is not sub-additive is not a good risk metric. The incentive for holding portfolios just dissolves if it is better to assess risk on individual positions and simply add up the total! When VaR is modelled using normal distributions for the risk factor returns VaR is always sub-additive. However, when the VaR model is extended to allow for heavy-tailed risk factor distributions then sub-additivity, in theory, need not hold. However, in practice market VaR is almost always found to be sub-additive; by contrast, there can be real problems when VaR is applied to credit risk, due to lack of sub-additivity.

Extreme value distributions, as their name suggests, examine the distribution of the extreme values of a random variable, which is typically assumed to be i.i.d. These extreme returns (or exceptional losses) are extracted from the data and an extreme value distribution may be fitted to these values. There are two approaches. Either one models the maximal and minimal values in a sample using the *generalised extreme value distribution* (GEV) or one models the excesses over a pre-defined threshold using the *generalised Pareto distribution* (GPD).

For instance, suppose the underlying time series consists of hourly observations on a portfolio P&L. On each day, we record the maximal loss and subsequently fit these daily data using a GEV distribution. Alternatively, we forget about the time spacing of the data, but record only those P&Ls that exceed a certain loss threshold, regardless of when this happens. Then we would fit the data using the GPD.

EVT models are often fitted using some form of maximum likelihood technique so the data requirements can be substantial. Very large data sets are crucial for EVT models as we are concerned only with the tail of the distribution and we require many extreme data points for robust estimation of parameters. The assumption that the observations are i.i.d. over such long sample periods is not very realistic in financial markets except, perhaps, when they have already been standardised using a volatility clustering model, as explained above. Nevertheless many banks do apply GEV distributions for intra-day VaR estimates. The GPD may be used to estimate conditional VaR; in fact there is a simple formula for this once the density has been fitted. However, the conditional VaR is quite sensitive to the choice of threshold loss, which must first be defined. For instance, the threshold could be set at the VaR that is estimated using a standard VaR model and then the GPD density can be fitted to obtain a more precise estimate of the conditional VaR.

III.A.3.5.3 VaR with Normal Mixtures

A normal mixture density function is a sum of normal density functions. For example, a mixture of only two normal densities $f_1(x) = \varphi(x; \mu_1, \sigma_1^2)$ and $f_2(x) = \varphi(x; \mu_2, \sigma_2^2)$ is the density function:²¹

$$g(x) = pf_1(x) + (1 - p)f_2(x).$$

The parameter p can be thought of as the probability that observation x is governed by density $f_1(x)$. In effect there are two *regimes* for x , one where x has mean μ_1 and variance σ_1^2 and another

²¹That is, $g(x) = p \left[(2\pi\sigma_1^2)^{-1/2} \exp\left(-\frac{1}{2}(x - \mu_1)^2 / \sigma_1^2\right) \right] + (1 - p) \left[(2\pi\sigma_2^2)^{-1/2} \exp\left(-\frac{1}{2}(x - \mu_2)^2 / \sigma_2^2\right) \right]$. Note that there is only one random variable so it would be misleading to call the densities f_1 and f_2 ‘independent’.

where x has mean μ_2 and variance σ_2^2 . In the case where x denotes the return on a portfolio, one can naturally identify these two regimes as a ‘high volatility’ (or even, for an equity portfolio, a ‘crash market’) regime with a low probability and the rest of the time a regime that governs ordinary, everyday market circumstances.

Consider a mixture of two zero-mean normal components, i.e. $\mu_1 = \mu_2 = 0$. In this case the variance is just

$$\text{NM}(2) \text{ variance} = p \sigma_1^2 + (1 - p) \sigma_2^2. \quad (\text{III.A.3.7})$$

The skewness is zero and the kurtosis is given by:

$$\text{NM}(2) \text{ kurtosis} = 3 \frac{p\sigma_1^4 + (1-p)\sigma_2^4}{[p\sigma_1^2 + (1-p)\sigma_2^2]^2}. \quad (\text{III.A.3.8})$$

The kurtosis is always greater than 3, so the mixture of two zero-mean normal densities always has a higher peak and heavier tails than the normal density of the same variance. For instance, Figure III.A.3.5 shows four densities:

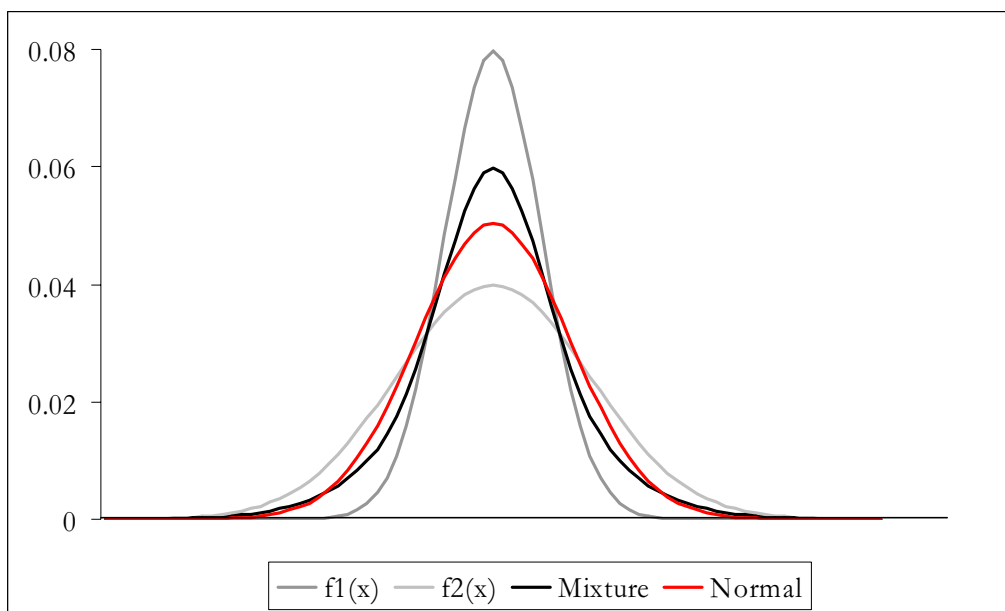
- three zero-mean normal densities with volatility 5%, 10% (shown in grey) and 7.906% (shown in red); and
- a normal mixture density, shown in black, which is a mixture of the first two normal densities with probability weight of 0.5 on each of the grey normal densities.

From formula (III.A.3.7), the variance of this mixture is $0.5 \times 5^2 + 0.5 \times 10^2 = 62.5$. Since $7.906 = \sqrt{62.5}$, the mixture has the same volatility as the red normal curve. However, it has kurtosis of 4.87 (substitute $p = 0.5$, $\sigma_1 = 5$ and $\sigma_2 = 10$ into formula (III.A.3.7)). In other words, it has an excess kurtosis of 1.87, which is significantly greater than zero, the excess kurtosis of the ‘equivalent’ normal (red) curve.

More generally, taking several components of different means and variances in the mixture can lead to almost any shape for the density. Maclachlan and Peel (2000) provide pictures of many interesting examples. The parameters of a normal mixture density function can be estimated using historical data. In this case the best approach is to employ the expectation–maximisation (EM) algorithm. The idea, as always, is to choose the parameters to maximise the likelihood of the data. But the EM algorithm differs from standard MLE in that the EM algorithm allows for some ‘hidden’ variables in the data that we cannot observe, so that we can only maximise the

expected value of the likelihood function and not the likelihood function itself.²² Alternatively, the parameters of a simple (e.g. two-component) normal mixture can be chosen in a scenario analysis of portfolio risk as, for instance, in Example III.A.3.7 below.

Figure III.A.3.5: A normal mixture density



There is no explicit formula for estimating VaR under the assumption that portfolio returns (or, equivalently, $P\mathcal{L}$) follow a normal mixture density. However, there is an implicit formula, so the problem is akin to that of implying volatility from the market price of an option (see Section II.G.1). That is, we can apply the Excel ‘Goal Seek’ (see Section II.G.1.4) or ‘Solver’ (see Section II.G.2.4) methods to back out the normal mixture VaR. To see how, suppose for the moment that we have a normal distribution for the $P\mathcal{L}$ of a linear portfolio. Then the analytic formula for VaR follows directly from the definition of VaR. That is, by definition,

$$\text{Prob}(P\mathcal{L} < -\text{VaR}_\alpha) = \alpha.$$

So if $P\mathcal{L}$ has a normal distribution with mean μ and standard deviation σ , we have

$$\text{Prob}(Z < [-\text{VaR}_\alpha - \mu]/\sigma) = \alpha,$$

where Z is a standard normal variable. Hence $[-\text{VaR}_\alpha - \mu]/\sigma = -Z_\alpha$, the α critical value of Z , and rearranging this gives our analytic formula for normal VaR:

$$\text{VaR}_\alpha = Z_\alpha \sigma - \mu. \tag{III.A.3.9}$$

²² The EM algorithm is far beyond the scope of the PRM syllabus, but interested readers should consult the book by Maclachlan and Peel (2000) which deals almost exclusively with this approach.

Using exactly the same type of argument we can derive the normal mixture VaR, but this time it is given by an implicit formula. For instance, with only two zero-mean components in the mixture (so there are only three parameters p , σ_1 and σ_2 , the standard deviations now being those of b -day P&L in each regime) we know everything *except* VaR_α in the identity:

$$p \text{Prob}(Z < -\text{VaR}_\alpha/\sigma_1) + (1 - p)\text{Prob}(Z < -\text{VaR}_\alpha/\sigma_2) = \alpha. \quad (\text{III.A.3.10})$$

Hence the b -day VaR_α can be ‘backed out’ from (III.A.3.10) using an iterative approximation method such as Goal Seek or Solver.

Example III.A.3.6: Scenario VaR using normal mixtures

A risk manager assumes there is a small chance, say 1 in 100, that the market will crash, in which case the expected portfolio return over a 10-day period is -50% with an (annualised) volatility around this mean return of 100% . However, at the moment in ordinary market circumstances there are steady positive returns of 10% per annum with a volatility of 20% . His portfolio is currently valued at $\$2\text{m}$. What is his 10-day VaR?

To answer this we extend (III.A.3.10) to the non-zero-mean case, and rephrase it in terms of the means and standard deviations of returns in the two regimes, rather than P&L. This gives:

$$p \text{Prob}(Z < [-\text{VaR}_\alpha - \mu_1]/P\sigma_1) + (1 - p)\text{Prob}(Z < [-\text{VaR}_\alpha - \mu_2]/P\sigma_2) = \alpha, \quad (\text{III.A.3.11})$$

where p is the probability of regime 1 (i.e. $1/100 = 0.01$), μ_1 is the 10-day return in regime 1 (i.e. -0.5), $\sigma_1 = 10$ -day standard deviation in regime 1 (i.e. $1/\sqrt{25} = 0.2$), μ_2 is the 10-day return in regime 2 (i.e. $0.1/25 = 0.004$), σ_2 is the 10-day standard deviation in regime 2 (i.e. $0.2/\sqrt{25} = 0.04$) and P is the current portfolio value ($\$2\text{m}$).

Using the Excel spreadsheet [NMVaR.xls](#) with Solver (or Goal Seek)²³ applied to cell B21 each time we change the significance level, we obtain the ‘NM VaR’ figures in the first row of the Table III.A.3.5.

Table III.A.3.5: NM VaR vs. normal VaR

Significance level	10%	5%	1%
NM VaR (\$)	157,480	500,003	1,012,632
<i>Equivalent</i> normal VaR (\$)	281,845	335,437	435,965
<i>Ordinary</i> normal VaR (\$)	94,524	123,588	178,107

²³ To apply Goal Seek, click Tools – Goal Seek – ‘Set cell B21 to value 0 by changing cell B18’.

The two ‘normal VaR’ figures are calculated using equation (III.A.3.9), or equivalently

$$\text{VaR}_\alpha = [Z_\alpha \sigma - \mu]P,$$

where μ and σ are the returns standard deviation and mean over the holding period. The ‘ordinary’ normal VaR figures are computed using the second (more likely) distribution of 10-day mean and standard deviation of:

$$\begin{aligned}\mu &= 0.004 \text{ (i.e. a 10\% annual return),} \\ \sigma &= 0.04 \text{ (i.e. a 20\% annual volatility).}\end{aligned}$$

That is, we ignore the possibility of a market crash in the ‘ordinary’ normal VaR. For the ‘equivalent’ normal VaR we use (III.A.3.7) to obtain an ‘equivalent’ standard deviation – and similarly the equivalent mean is $p\mu_1 + (1 - p)\mu_2$.

These adjust the ‘ordinary’ market circumstances mean and standard deviation to take account of the possibility of a crash, but after that the VaR is computed using the normal assumption for portfolio returns.

From the results in Table III.A.3.5 it is clear that ignoring the possibility of a crash can seriously underestimate the VaR. Even if one were always to assume a normal distribution, the VaR will be almost three times larger when the standard deviation and mean are adjusted to account for the possibility of a crash. It is the relationship between the NM VaR and the ‘equivalent’ normal VaR that is really interesting. Our example exhibits some typical features of NM VaR:

- For low significance levels (e.g. 10% or 20%), the normal assumption can seriously *overestimate* VaR (in this example, it was about twice the size of the NM VaR).
- For higher significance levels (e.g. 5% or 1%), the normal assumption can seriously *underestimate* VaR (in this example, it was about half the size of the NM VaR).

The significance level at which the NM VaR becomes greater than the normal VaR based on an equivalent volatility/mean depends on the degree of excess kurtosis in the data. When the excess kurtosis is relatively small it may be that the 5% (or even 1%) VaR is actually *smaller* under the NM assumption. In fact, when the parameters are estimated using actual historical observations on the portfolio returns it is common to find that the normal mixture VaR is less than the normal VaR at the 5%, and even at the 1% level.

Finally, it should be noted that, although we have not described the generalisation of the MC VaR model to the normal mixture case, this is a simple (one- or two-line) extension to the simulation code. We outline the method for the two-component zero-mean case:

- Define two risk factor covariance matrices \mathbf{V}_1 and \mathbf{V}_2 and associated probability weights ($p, 1 - p$) in the mixture (either estimated from historical data using the EM algorithm, or assumed in a scenario analysis).
- Break each of the 10,000 or so simulations into two steps: (a) draw from a Bernoulli variable with probability p ; (b) if the result is ‘success’ use \mathbf{V}_1 to generate the correlated risk factors in the simulation, else use \mathbf{V}_2 .

In summary, the normality assumption is not necessary for the analytic and Monte Carlo VaR methodologies. This section has shown how the analytic and MC VaR methods can be used with a normal *mixture* distribution for the portfolio returns. Such a distribution is better able to capture the skewness and excess kurtosis that we commonly observe in portfolios of most types of financial assets. Hence, if the parameters of the normal mixture distribution are estimated from historical data, the resulting VaR estimate will reflect the actual properties of the data more accurately than the normal VaR. Another very useful application of normal mixture VaR is to probabilistic scenario analysis, where the portfolio returns are generated by a high-volatility (or ‘crash’) component with a low probability and, the rest of the time, by another component that applies to the ordinary market circumstances.

III.A.3.6 Decomposition of VaR

As explained in Chapter III.0, VaR models form the basis of internal economic capital allocation and limit setting. Hence, firms need to aggregate VaR over different risk types and over different business activities and, likewise, to disaggregate VaR into different components. Disaggregation of risk is used in risk management for setting limits, assessing new investments, hedging and performance measurement. It allows risk managers to understand the drivers of risk in their portfolio.

A number of different rules for disaggregating risk are considered in this section. Their common theme is that each rule is based on the analytic VaR formula; that is, the rule is derived using the rules for the variance operator (see Section II.E.3.4). However, it is common practice, for instance in the allocation of economic capital, to apply these rules to all VaR estimates, irrespective of the model used to compute them. But not all these ‘rules’ for VaR decomposition are appropriate for historical or MC VaR estimates. Here the VaR corresponds to a percentile, but percentiles do not satisfy simple rules, like the variance operator. Hence the usefulness of each rule varies, depending on the intended application.

III.A.3.6.1 Stand-alone Capital

Suppose a line manager operating on a VaR-based risk limit wants to assign separate VaR limits to the equity and foreign exchange desks so that aggregate losses only exceed the aggregate VaR limit an appropriately small proportion of the time. But since risk limits do not correspond to real capital, it is not necessary for the two limits to add up to his overall VaR limit. On the contrary, in theory it could even be that the VaR limit for the equity desk, say, *exceeds* the overall risk limit! We shall see why, and how, in this section.

In Example III.A.3.3 we considered a simple portfolio that has been mapped to two risk factors, an equity index and an exchange rate. The total 1% 10-day VaR due to both risk factors was estimated as \$319,643. However, the VaR due to equity risk alone was:

$$\begin{aligned}\text{Equity VaR} &= 2.33 \times 10\text{-day standard deviation FTSE} \times \$3\text{m} \\ &= 2.33 \times 0.15 \times \sqrt{10/250} \times 3 = \$209,700.\end{aligned}$$

Similarly,

$$\text{FX VaR} = 2.33 \times 0.2 \times \sqrt{10/250} \times 2 = \$186,400.$$

Hence,

$$\text{Equity VaR} + \text{FX VaR} = \$396,100 > \$319,643 = \text{Total VaR}.$$

In the normal analytic VaR model VaR follows the same ‘rules’ as the standard deviation operator.²⁴ In this case it is easy to show that VaR is ‘sub-additive’ in the sense that:

$$\text{Total VaR} \leq \text{Sum of component VaRs}$$

with equality if, and only if, all the correlations in the covariance matrix \mathbf{V} are one.

We shall state the complete rule for the decomposition into *two* components in (II.A.3.12) below. It shows that the total VaR is only equal to the sum of the component VaRs if there is perfect correlation between the components. But in the above example the risk factors had a correlation of 0.3, which is much less than one. So the total VaR was much less than the sum of the two component VaRs. In fact, had the correlation been large and negative, the total VaR might actually have been less than *either* of the component VaRs.

The fact that VaR aggregates take account of correlations, and that these are typically far less than one, is one of the reasons why banks favour VaR over ‘traditional’ risk measures, such as duration for a bond portfolio, or the Greeks for an options portfolio, or beta for an equity portfolio. These traditional risk measures ignore the benefits of diversification that apply in a

²⁴ But only when the risk factor returns are assumed to be normal: as mentioned already in Section III.A.3.5.2, if risk factor returns are heavy-tailed then VaR need not be sub-additive.

portfolio exposed to multiple risk factors. In contrast, VaR accounts not only for the risks due to the risk factors themselves, but also for the less than perfect correlation of risk factors when aggregating the risk. Also, VaR is a risk measure with consistent dimensions across markets and therefore allows greater consistency in setting policy across products and in evaluating the relationship of risk and return.

III.A.3.6.1.1 Decomposing non-linear portfolios

Decomposition of risk becomes more complex for portfolios with non-linearities. VaR measures for such portfolios are much more likely to violate the criterion of sub-additivity. For this reason VaR is criticised as being a poor 'risk metric'. It is possible to construct extreme cases where individual portfolios containing, say, short, well out-of-the-money digital options have a very low or even zero VaR at, say, the 95th percentile. However, if two similar portfolios are combined together, the diversified portfolio has higher VaR than each of the components. In this anomalous situation, stand-alone capital is not appropriate for limit setting.²⁵ Indeed, a strong argument could be made here to avoid VaR and to use conditional VaR instead, for limit-setting purposes.

Of course typically, non-linear portfolios will be analysed using the simulation approaches. Disaggregation of VaR (or conditional VaR) can be undertaken in the context of simulation by restricting the risk factor scenarios in different ways. For instance, total VaR can be disaggregated into an equity VaR component, corresponding to a lower percentile of a simulated portfolio returns distribution with only the equity risk factors changing, and an FX VaR component, corresponding to a lower percentile of a simulated portfolio returns distribution with only the foreign exchange rates changing. Since percentiles do not obey simple 'rules' (except that a percentile is invariant under a monotonic transformation of variables) there is, in this case, no simple rule that relates the sum of equity VaR and FX VaR to the total VaR.

III.A.3.6.1.2 Specific vs. Systematic Risk

Another way of disaggregating risk is to decompose VaR into its systematic and specific components. That is, those risks that apply to the market/factor generally and those that arise from lack of diversification or deviations from the market portfolio. A VaR model can be used to assess the 'specific risks' of a portfolio – i.e. the risks that are not captured by the risk factor mapping.

²⁵ Each separate portfolio could be within limit yet the business overall could be in breach of limits when considered on a diversified basis.

Example III.A.3.7: Specific VaR

To obtain the beta of 1.5 for our portfolio of UK equities in Example III.A.3.3 we re-created an artificial price history of the portfolio using the current portfolio weights, and regressed the time series of returns to this portfolio on the FTSE index returns. This gave the Excel output shown in Table III.A.3.6.

Table III.A.3.6: A CAPM regression

Regression Statistics		Coefficients		Standard Error	t Stat
R Square	0.7284	Intercept	-0.00016	0.00045	-0.34991
Standard Error	0.00802	FTSE	1.50312	0.02876	52.26426

The ‘standard error’ is the standard deviation of the model residuals (see Section II.F.6). Since the returns are observed daily, the 10-day standard deviation of the residuals in this model is

$$0.00802 \times \sqrt{10} = 0.02536$$

Hence, the 1% 10-day specific VaR is $2.33 \times 0.02536 \times \$2m = \$118,184$.

Of course, having re-created an artificial price history of the portfolio using the current portfolio weights, we did not *need* to estimate a factor model in order to calculate the total VaR. We could have simply obtained the daily standard deviation of the ‘re-created’ portfolio returns and used formula (III.A.3.4). In this case, we would have a daily standard deviation 0.01636, giving a direct estimate of 1% 10-day total VaR as:

$$\text{Total VaR} = 2.33 \times 0.01636 \times \sqrt{10} \times \$2m = \$241,117.$$

However, the 1% 10-day systematic VaR (in this case, the equity risk factor VaR) is \$209,700. Adding this to the specific VaR of \$118,184 we thus obtain:

$$\text{Systematic VaR} + \text{Specific VaR} = \$327,884.$$

Clearly, simply adding systematic VaR and specific VaR is a very conservative way to estimate the total VaR, that is, direct estimation of total VaR will normally give a result that is *considerably* lower than the sum of systematic and specific VaR.

III.A.3.6.1.3 Sub-additivity

To understand why this is so, we do some simple algebra showing that the analytic VaR (for linear positions) obeys the following *sub-additive* rule for any decomposition of total VaR into two components, VaR₁ and VaR₂ where the component risks have correlation ρ:

$$\text{Total VaR}^2 = \text{VaR}_1^2 + \text{VaR}_2^2 + 2 \rho \text{VaR}_1 \text{VaR}_2. \tag{III.A.3.12}$$

Note that expression (III.A.3.12) simplifies when the correlation is one or zero:

$$\begin{aligned} \text{If } \rho = 1: \quad & \text{Total VaR}^2 = \text{VaR}_1^2 + \text{VaR}_2^2 + 2 \text{VaR}_1 \text{VaR}_2 = (\text{VaR}_1 + \text{VaR}_2)^2 \text{ and} \\ & \text{Total VaR} = \text{VaR}_1 + \text{VaR}_2. \\ \text{If } \rho = 0: \quad & \text{Total VaR}^2 = \text{VaR}_1^2 + \text{VaR}_2^2 \text{ and} \\ & \text{Total VaR} = \sqrt{(\text{VaR}_1^2 + \text{VaR}_2^2)}. \end{aligned} \tag{III.A.3.13}$$

Now recall that if the factor model is capturing most of the variation in the portfolio then specific risks and systematic risks should be *uncorrelated* (see Section II.F.2). In that case, the Total VaR will be closer to the square root of the sum of the squared component VaRs than to the simple sum of the two VaRs. In other words, simply *adding* specific VaR and systematic VaR is not a good way to estimate Total VaR if the systematic and specific components are (more or less) uncorrelated.

When disaggregating VaR into different components, it is sometimes assumed that all component correlations lie between zero and one. In this case the two ways of calculating total VaR (as a straightforward sum of VaRs, or as the square root of the sum of the squared VaRs) provide approximate *upper* and *lower* bounds for the total VaR, respectively.

III.A.3.6.2 Incremental VaR

Incremental VaR (IVaR) is a measure of how portfolio risk changes if the portfolio itself is changed in some way. It is ideal for assessing the effect of a hedge or a new investment decision on a trader's VaR limit, or for assessing how total business risk would be affected by the sale/purchase of a business unit. There are two ways of proceeding:

- (a) *The before and after approach.* Here we measure the VaR under the proposed change, compare it to the current VaR and take the difference. This is the best approach if the proposed change to the portfolio is significant.
- (b) *The approximation approach.* We can find an approximate IVaR by first calculating the *DeIVaR* vector (see below). This vector is then multiplied by another vector containing the proposed changes in positions for each asset/risk factor. Like all approximations based on partial derivatives, it is suitable only for examining small changes to the portfolio composition. Under the analytical method²⁶ the DelVar vector is calculated as follows:

²⁶ With simulation methods a DelVar vector could also be constructed, but with considerably greater difficulty. Each element of the vector would be determined by assessing the change in total VaR (according to simulation) for a one-unit change in the portfolio holdings of the relevant asset.

$$DelVaR = \frac{\mathbf{V}\mathbf{p}Z_{\alpha}}{(\mathbf{p}'\mathbf{V}\mathbf{p})^{0.5}}, \quad (\text{III.A.3.14})$$

where \mathbf{V} is the covariance matrix, \mathbf{p} is the position vector and Z_{α} is the relevant critical value of the normal distribution. Note that the denominator of this expression is simply the standard deviation of the portfolio's P&L. The DelVaR vector will contain an element for each asset/risk factor in the covariance matrix (the first element will contain information relating to the risk of the first asset/risk factor, and so forth).

Example III.A.3.8 Approximate IVaR

We continue with the example first presented in Example III.A.3.3. The portfolio consists of an exposure to the FTSE index and exposure to the exchange rate since the investor is US\$ based. Note that the standard deviation of the portfolio's P&L over a one-day horizon is equal to \$0.043382m (being $\sqrt{0.001882}$). Hence, for the base position where the portfolio is unchanged, equation (III.A.3.14) becomes:

$$\mathbf{V}\mathbf{p} = \begin{pmatrix} 0.00009 & 0.000036 \\ 0.000036 & 0.00016 \end{pmatrix} \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 0.000342 \\ 0.000428 \end{pmatrix},$$

$$DelVaR = \begin{pmatrix} 0.000342 \times 2.33 \div 0.043382 \\ 0.000428 \times 2.33 \div 0.043382 \end{pmatrix} = \begin{pmatrix} 0.018368 \\ 0.022987 \end{pmatrix}.$$

Now suppose that we consider hedging half of the currency exposure so that the exposure to the exchange rate is reduced to only \$1m, rather than the current \$2m. We can construct a vector of portfolio changes where the first element (the change in the exposure to FTSE) is equal to zero, and the second element (the change in the exposure to currency) is -1 . The incremental VaR can be approximated as follows:

$$IVaR = \begin{pmatrix} 0 & -1 \end{pmatrix} \times \begin{pmatrix} 0.018368 \\ 0.022987 \end{pmatrix} = -0.022987.$$

In other words, the hedging decision will reduce 1% one-day VaR by approximately \$22,987.

As the proposed change to the portfolio is quite large in this case, the approximation method is unlikely to be accurate and the before and after method would be preferable. With the before and after method, we know from Example III.A.3.3 that the one-day VaR of the original position is

$$2.33 \times \sqrt{0.001882} \text{ m\$} = \$101,080.$$

With the currency hedge, the 1% one-day VaR is \$80,241 because

$$\mathbf{p}'\mathbf{V}\mathbf{p} = (3 \quad 1) \begin{pmatrix} 90 & 36 \\ 36 & 160 \end{pmatrix} \begin{pmatrix} 3 \\ 1 \end{pmatrix} \times 10^{-6} = (306 \quad 268) \begin{pmatrix} 3 \\ 1 \end{pmatrix} \times 10^{-6} = 0.001186,$$

and $2.33 \times \sqrt{0.001186} \text{ m\$} = \$0.08241\text{m}$. Hence, according to the exact method, the IVaR is $101,080 - 80,241 = \$20,839$, which is *less* than the IVaR given by the approximation. In other cases the actual IVaR will exceed the approximation, depending on the level of correlation between assets.

III.A.3.6.3 Marginal Capital

Sometimes referred to as *component VaR* (CVaR), this method of decomposition is useful for gaining a better understanding the drivers of risk within a portfolio. Sometimes it is also used for performance measurement. Unlike the previous methods, marginal capital is additive; the sum of each of the component VaRs will be equal to the total VaR.²⁷

We take the DelVaR vector discussed above and multiply each element by the corresponding position for each asset/risk factor. The result is a set of CVaRs for each asset. The sum of these CVaRs will be equal to portfolio VaR. Since the analysis is performed using the DelVaR vector of partial derivatives with respect to asset weights, it is only relevant for the current portfolio weightings. Significant changes to the portfolio will change the risk contributions of the various assets, necessitating new analysis based on the revised DelVaR vector.

Example III.A.3.9: Marginal capital

We continue to analyse the portfolio first introduced in Example III.A.3.3. The DelVaR vector, calculated in Example III.A.3.8 is used here:

$$DelVaR = \begin{pmatrix} 0.018368 \\ 0.022987 \end{pmatrix}.$$

The CVaR for equities is equal to the position (\$3m) multiplied by 0.018368, or \$55,104. The CVaR for foreign exchange is equal to the position (\$2m) multiplied by 0.022987, or \$45,974. Note that they sum to \$101,078, which is approximately equal to \$101,080, the 1% one-day VaR (the error is due to rounding in the DelVaR vector).

CVaR is used to assign a *proportion* of the total risk that can be attributed to each component. For instance, in this example the equity exposure is contributing $55,105/101,078$ or around 55% of

²⁷ This relationship is assured for the analytical method and if the portfolio is comprised of simple linear positions. It may not hold for other cases.

the risk in the portfolio, while currency contributes around 45%. This technique can be used to evaluate the risk of an asset (or asset class) in the context of a *diversified* portfolio. In contrast, the stand-alone capital (in Section III.A.3.6.1) measures the risk of an asset class in isolation. Either can be used for performance measurement purposes, although stand-alone capital is most commonly used as the risk measure in a risk-adjusted performance measure context.

For instance, we would use stand-alone capital to compare the performance of the equity and foreign exchange trading desks because it is generally argued that the diversification benefit should not enter into the analysis. That is, the team managing foreign exchange risks presumably has no say in the way the portfolio of businesses is constructed (whether or not there is an equity desk, and the relative sizes of those businesses). They should therefore be neither rewarded nor penalised for the diversification of the overall portfolio of businesses. Instead, performance measures should be centred only on the issues over which they have direct control, being foreign exchange risks in this example.

III.A.3.7 Principal Component Analysis

Principal component analysis (PCA) is a statistical tool that decomposes a positive semi-definite matrix into its *principal* components.²⁸ For instance, Section II.D.5.5 shows how PCA on an $n \times n$ covariance matrix is used to write the portfolio variance as a sum of n positive terms that become progressively smaller, with the first principal component explaining the largest part of the variation in the system represented by the covariance matrix and so on. The m th principal component explains the least variation – indeed, the variation captured by the lower-order principal components is commonly ignored, because it just picks up the ‘noisy’ variation that we would prefer to ignore.

PCA applied to a covariance matrix or a correlation matrix has many applications to financial risk management. It is particularly effective in highly correlated systems such as term structures, i.e. yield curves, or futures prices or even implied volatilities. Here only few components are needed to explain almost all of the variation. In this respect PCA is a useful technique for *reducing dimensions*. By retaining only the first few principal components – enough, say, to explain 95% of the variation observed historically – it cuts out the ‘noise’ for the subsequent analysis.

The other great advantage of PCA is that the principal components are *uncorrelated with each other*. This means, for instance, that one can perform a simple scenario analysis on each of the main principal components separately, leaving the other components fixed. The change translates into

²⁸ Positive definiteness is discussed in detail in Section II.D.5.4.

a meaningful scenario, i.e. one that could be observed historically. For instance, when applying PCA to a covariance matrix of zero-coupon yields of different maturities,²⁹ a scenario for the change in the first principal component generally will mimic an almost parallel shift in the entire yield curve.

Without PCA one would need to take care that only ‘correlated’ scenarios are used – for instance, if the scenario specifies that the one-month interest rate increases by 100 bps, one could not have the three-month interest rate decreasing by 200 bps in the same yield curve scenario. Correlated scenarios are generated using the Cholesky decomposition of a covariance matrix – see Section II.D.4.2. This problem is compounded if simulations are extended more than a short time into the future using this method, as it is difficult to prevent generating implausible shapes in the resulting yield curves.

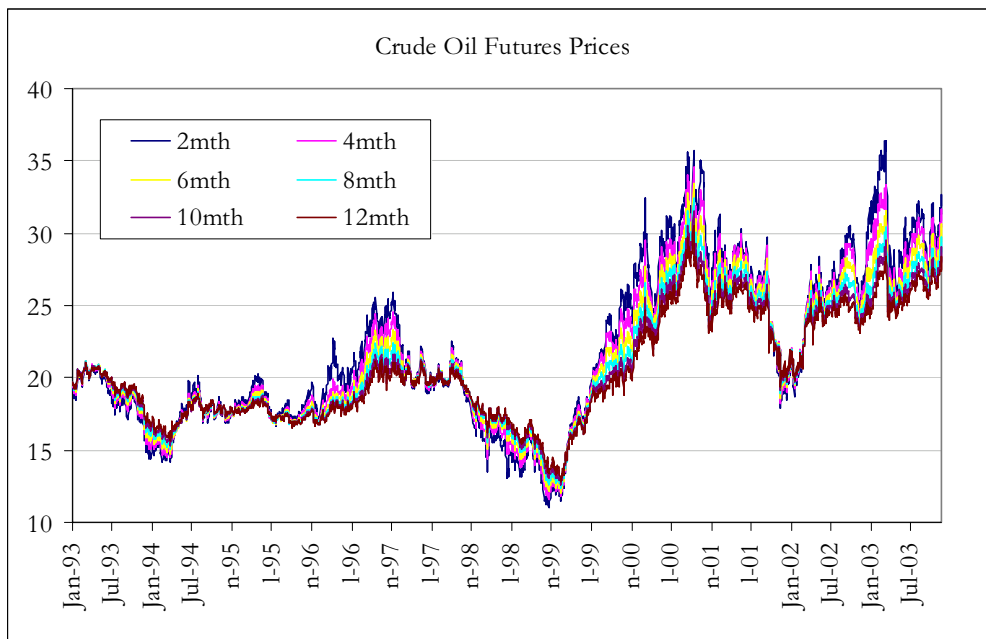
III.A.3.7.1 PCA in Action

From 4 January 1993 until 20 November 2003 we have daily closing New York Mercantile Exchange (NYMEX) futures prices, from 2 to 12 months out, on West Texas Intermediate (WTI) light sweet crude oil and on natural gas. Some of these are shown in Figure III.A.3.6. The system of futures returns is clearly very highly correlated – indeed there are perhaps just one or two independent sources of information driving the whole system of futures prices.

The correlation matrix of the returns to futures from 2 to 12 months out, based on the entire sample from 4 January 1993 until 20 November 2003, is shown in Table III.A.3.7. This exhibits the pattern that is typical of term structures, with correlation decreasing as the maturity difference increases. The correlations are so high that almost all the variation can be attributed to two or perhaps three components. An analysis of eigenvalues tells us how many components we need. The eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ of an $n \times n$ correlation matrix sum to n , and the amount of variation captured by the i th principal component is λ_i/n . So in our example, with $n = 11$, if the largest eigenvalue is, say, 10, then the first principal component explains $10/11 = 90.9\%$ of the variation.

²⁹ More precisely, to the covariance matrix of the daily (or weekly or monthly) *changes* in yields.

Figure III.A.3.6: A highly collinear system



	<i>m2</i>	<i>m3</i>	<i>m4</i>	<i>m5</i>	<i>m6</i>	<i>m7</i>	<i>m8</i>	<i>m9</i>	<i>m10</i>	<i>m11</i>	<i>m12</i>
<i>m2</i>	1										
<i>m3</i>	0.993	1									
<i>m4</i>	0.984	0.997	1								
<i>m5</i>	0.974	0.990	0.998	1							
<i>m6</i>	0.963	0.981	0.993	0.998	1						
<i>m7</i>	0.951	0.972	0.986	0.994	0.998	1					
<i>m8</i>	0.939	0.962	0.978	0.988	0.995	0.999	1				
<i>m9</i>	0.927	0.951	0.969	0.981	0.990	0.996	0.999	1			
<i>m10</i>	0.915	0.940	0.960	0.974	0.984	0.991	0.996	0.999	1		
<i>m11</i>	0.901	0.928	0.949	0.965	0.977	0.986	0.992	0.996	0.999	1	
<i>m12</i>	0.888	0.916	0.938	0.956	0.969	0.979	0.987	0.992	0.996	0.999	1

Table III.A.3.8 shows the eigenvalues and the first three eigenvectors of this matrix (see Section II.D.5 for an explanation of these). The first three eigenvalues show that the first principal component explains 97.5% of the total variation, the second component explains a further 2.2% (since $0.245/11 = 0.022$) and the third component only explains a tiny amount, 0.145% of the movements in the futures prices in our sample.

Table III.A.3.8: Eigenvectors and eigenvalues

<i>Eigenvalues</i>	<i>Future</i>	<i>1st Eigenvector</i>	<i>2nd Eigenvector</i>	<i>3rd Eigenvector</i>
10.732	<i>m2</i>	0.293	0.537	0.623
0.245	<i>m3</i>	0.299	0.412	0.078
0.016	<i>m4</i>	0.302	0.280	-0.225
0.004	<i>m5</i>	0.304	0.161	-0.348
0.002	<i>m6</i>	0.305	0.054	-0.353
0.001	<i>m7</i>	0.305	-0.044	-0.276
0.000	<i>m8</i>	0.304	-0.132	-0.162
0.000	<i>m9</i>	0.303	-0.211	-0.028
0.000	<i>m10</i>	0.302	-0.282	0.115
0.000	<i>m11</i>	0.300	-0.350	0.247
0.000	<i>m12</i>	0.298	-0.411	0.364

The first three eigenvectors in Table III.A.3.8 are used to compute the first three principal components as:

$$PC1 = 0.293 m2 + 0.299 m3 + \dots + 0.298 m12, \quad (\text{III.A.3.12a})$$

$$PC2 = 0.537 m2 + 0.412 m3 + \dots - 0.411 m12, \quad (\text{III.A.3.12b})$$

$$PC3 = 0.623 m2 + 0.078 m3 + \dots + 0.364 m12, \quad (\text{III.A.3.12c})$$

where $m2, \dots, m12$ denote the (normalised) returns to the futures of different maturities from 2 to 12 months out.

Since the eigenvectors are always orthogonal (see Section II.D.5.2), equations (III.A.3.12) can be rewritten as:

$$m2 = 0.293 PC1 + 0.537 PC2 + 0.623 PC3,$$

$$m3 = 0.299 PC1 + 0.412 PC2 + 0.078 PC3,$$

..

..

$$m12 = 0.298 PC1 - 0.411 PC2 + 0.364 PC3.$$

This is known as the *principal component representation* of the system. Since the coefficients on PC1 are all approximately the same, when PC1 moves (holding the other components fixed) the term structure of futures prices will shift in (almost) parallel fashion. For this reason PC1 is often called the ‘trend component’ when PCA is applied to term structures. When PC2 increases the term structure of futures prices will shift up at the short end but down at the long end – so PC2 is called the ‘tilt component’. And when PC3 increases the term structure shifts up at both ends but – looking again at Table III.A.3.8 – the medium term futures prices will go down. Hence PC3 is called the ‘curvature’ component.

In this example the curvature component is not important – it corresponds to much less than 1% of the movements normally found in futures prices. But in other examples – interest rates, for instance – the curvature component can be more important. And when PCA is applied to other types of systems, such as equities or currencies, we normally need more than three components to model the system with sufficient accuracy. Of course, there will be no intuitive interpretation of the components as there is for term structures – because we cannot ‘order’ a system of equities or currencies in any sensible way – although the first principal component will normally capture the common trend (if the system is sufficiently correlated that there *is* a common trend!).

III.A.3.7.2 VaR with PCA

PCA is commonly used in the analytic VaR method for cash flows, and the MC simulation VaR method for interest-rate options portfolios. Perhaps the most important input in both these approaches is the covariance matrix of risk factor returns – in this case, the covariance matrix of changes in a zero-coupon yield curve. Typically this yield curve will have more than 10 different maturities, so the dimension of the risk factor space is large. But, as we discussed earlier, large-dimensional covariance matrices are very difficult to estimate using GARCH models. And EWMA matrices normally assume the same value for the smoothing constant for *all* the risk factors. This may seem fine if the risk factors are just one yield curve, but for international fixed income portfolios the risk factors consist of many yield curves.

In order to take proper account of (less than) perfect correlation between these risk factors when estimating the total VaR, we need to use a covariance matrix of the whole system – i.e. all yields of all maturities in all countries. Then in the EWMA matrix it is very unrealistic to apply the same value for the smoothing constant for *all* the risk factors. And direct estimation of the matrix using a multivariate GARCH model will be out of the question because the likelihood surface will not be well defined.

The solution is to apply PCA to the entire system of risk factors, retaining enough components in the principal component representation to explain most (say, 95%) of the variation. Then, because they are uncorrelated, it is only the *variance* of each component that matters – their covariances are zero.³⁰ So one can treat each component separately, estimating and forecasting only its variance using either GARCH or EWMA. Note that with this method we do *not* have to use the same smoothing constant in each EWMA; and even if we do use the same smoothing constant, the final risk factor covariance matrix will not have the same effective smoothing constant for all risk factors, as happens in the RiskMetrics methodology. Nor do we need to

³⁰ Note that it is only the *unconditional* covariances that are zero, because PCA is based on an *unconditional* covariance matrix, which does not have time-varying covariances. Hence the application of a time-varying (i.e. EWMA or GARCH) model to variances, whilst assuming correlations are constant, does entail a strong assumption.

constrain the GARCH models in any particular way, as in most multivariate GARCH models. The large risk factor covariance matrix that we obtain in the end will always be positive semi-definite. Full details are given in Alexander (2001).

III.A.3.8 Summary

The crucial question for any VaR model is whether the VaR estimate provides a good indication of the ‘true’ risk in financial markets. Any deviation between the VaR estimate and this truth represents a model risk. Such a risk is potentially disastrous since it could cause a financial institution to have insufficient capital given the risks taken, making it more vulnerable than it should be to insolvency. Concern about this kind of model risk is well justified because financial markets are prone to periods of high volatility in which extreme market movements can occur. Many analysts have noted the heavy tails of empirical distributions relative to the standard i.i.d. normal assumptions. These are a symptom of the clusters of volatility that occur periodically.

Simple VaR models as explained in the previous chapter are all flawed to some extent; either their assumptions are too simple and/or they suffer from lack of data. This chapter has explored various methods for improving on VaR models so that they are more consistent with the behaviour observed empirically in financial markets.

We have argued that the most crucial issue for analysts to address in this regard is volatility clustering. It is now well established that volatility in financial markets exhibits elevated values over certain periods before reverting to lower levels. The variation in volatility explains most of the extreme moves observed in financial markets. We focused on models that incorporate the volatility clustering idea (EWMA and GARCH) and considered their practical application to VaR estimation. We also examined some alternative approaches to the issue of extreme returns that do not incorporate volatility clustering (Student’s t , EVT and normal mixtures).

Of course, it should always be remembered that model risk can never entirely be eliminated. The advanced models presented here have their own problems, and these should be well understood by the user. Every model is a simplification of the reality it represents, and should be used with caution. The pervasiveness of model risk is a key reason why stress tests (covered in the next chapter) are useful.

This chapter also completed the discussion of VaR for market risk by examining two other, more advanced topics: risk decomposition and principal component analysis. We have described the ways in which total VaR can be ‘disaggregated’ into different components, for capital allocation on a ‘stand-alone’ and on a ‘marginal’ basis, and how traders can use an incremental VaR analysis

to evaluate the effect of a proposed trade on their VaR limit. Finally, we have explained how PCA can greatly simplify VaR estimation when there are multiple risk factors that are highly correlated. For instance, the VaR for any portfolio of bonds or loans that are mapped to risk-equivalent cash flows (so that the risk factors are a term structure of interest rates) should be calculated using PCA. Similarly, commodity portfolios with exposures to futures prices at different maturities are best represented using VaR with PCA than by a direct VaR analysis. Not only does this simplify the VaR estimation itself, but the use of PCA greatly facilitates the application of stress tests and scenario analysis to these types of portfolios, as we shall see in the next chapter.

References

- Alexander, C (2001) *Market Models: A Guide to Financial Data Analysis*. Chichester: Wiley.
- Bauwens, L, Laurent, S, and Rombouts, J (2003) ‘Multivariate GARCH models: a survey’. To appear in *Journal of Applied Econometrics*. Available at <http://www.core.ucl.ac.be/econometrics/Bauwens/Papers/papers.htm>
- Berkowitz, J, and O’Brien, J (2002) ‘How accurate are value-at-risk models at commercial banks?’ *The Journal of Finance*, LVII, pp. 1093–1111.
- Bollerslev, T (1990) ‘Modelling the coherence in short-run nominal exchange rates: A multivariate Generalised ARCH model’, *Review of Economics and Statistics*, 72, pp. 498–505.
- Dowd, K (2002) *Measuring Market Risk*. Chichester: Wiley.
- Maclachlan, G, and Peel, D (2000) *Finite Mixture Models*. New York: Wiley.

III.A.4 Stress Testing

Barry Schachter¹

III.A.4.1 Introduction

The previous chapters have introduced value-at-risk (VaR) as a risk measure and discussed methods to deal with some of its shortcomings. It is common belief that VaR does not provide a complete picture of portfolio risk and that stress testing is a means of addressing that (at least in part). While most of this chapter is devoted to questions related to the construction of stress tests, it is important to step back and formulate some ideas about why we need to have a chapter on stress testing.

By now the need for stress testing of portfolios of financial instruments is taken for granted. We find stress testing on every list of risk management best practices. Stress testing has long since been solemnised by regulators. If this chapter were to be simply a review of stress-testing techniques, then I could immediately proceed by providing a typology and discussion of the various techniques that have been presented in the literature on the topic. I think, however, that an appreciation of the usefulness of stress testing requires that I first step back a bit and ask a couple of basic questions. The first question is by what *historical process* we have arrived at this general acceptance of the need for stress testing. The second question is by what *conceptual framework* we have concluded that stress testing is needed.

My definition of stress testing is as follows:

Stress testing (strēs' tĕst'ing) n.

1. *A method for the quantification of potential future extreme, adverse outcomes in a portfolio of financial instruments.*
2. *A palliative for the anxiety that is experienced by managers with significant risk exposures.*

The key words in this definition are *quantification*, *potential future*, *extreme* and *palliative*. I am taking a very broad view of stress testing here to include just about any method for attaching a monetary figure (i.e., a *quantity*) to potential losses from *extreme* events that is not specifically VaR. It is important to note that *quantitative* estimates are not necessarily statistical estimates. In fact, most stress-testing methods, unlike VaR, are not statistical measures of risk. That is, no probabilities are attached to and no confidence intervals estimated for the adverse outcomes. It is a great

¹ Chief Risk Officer, Balyasny Asset Management, LLC. Thanks go to Steve Allen and Carl Batlin for reviewing and commenting on a draft of this chapter. Any remaining errors or omissions are my own. Parts of this chapter draw on Schachter (1998a, 1998b, 2000a, 2000b).

challenge in stress testing, whatever the method employed, from historical scenario analysis (reliving a past market crisis) to factor push (shifting a market rate to see the impact on a portfolio), to effectively gauge *potential future* losses. A *palliative* is something that reduces pain, but does not cure an underlying problem. To a large extent, this is the role that stress tests have served to date, which is somewhat unsatisfying. The risk management profession is still seeking consensus on theoretically and practically consistent ways to integrate stress testing into decision making. Uses vary widely from a simple informational tool to a formal element of limit setting and capital allocation.

The last point above deserves emphasis. Even if we agree what stress testing is, we still need to agree (or at least understand) its purpose. Stress testing must fit into some wider context, or have some point. We should be able to establish how stress testing provides some incremental value as either a means to achieving some goal (e.g., the optimal allocation of assets in a portfolio) – in other words, it is a direct input in some decision function – or a way of measuring the movement towards some goal – in other words, it is a benchmark or feedback mechanism for modifying or improving a decision. If we cannot show either of these things, then we cannot know how to make rational economic decisions which use stress-test results, and we cannot articulate why we should be performing stress tests.

It is common belief that stress testing has incremental value because VaR is a sufficient statistic for risk under only a restricted set of conditions. For instance, in a world where all financial asset returns are jointly normally distributed and portfolios are passive, the magnitude of a loss of any probability is a scalar multiple of the VaR. The loss on a portfolio corresponding to a one standard deviation move in the portfolio value is equal to the 99% VaR divided by 2.33. To know the VaR is to know everything about risk, and stress tests are not needed. However, if asset returns are nonlinear functions of underlying market risks, or market risks are not distributed according to some distribution that is stable under addition, then to know VaR is to have only an incomplete picture of portfolio risk. Stating it more colloquially, while VaR models provide a notion of a ‘bad’ portfolio return, they do not convey just how bad ‘bad’ can get. Even if we agree that stress testing has incremental value, then the context for extracting that value in decision making still needs to be shown. This chapter, focusing mainly on methods for stress testing, does not settle this issue.

III.A.4.2 Historical Context

In the classic 1967 film, *The Graduate*, Benjamin Braddock (Dustin Hoffman) returns home after graduation from university, and receives the following sage (and frightening) career advice from a family friend:

MR MCGUIRE: I just want to say one word to you. Just one word.

BEN: Yes sir.

MR MCGUIRE: Are you listening?

BEN: Yes I am.

MR MCGUIRE: Plastics.

BEN: Exactly how do you mean?

MR MCGUIRE: There's a great future in plastics. Think about it. Will you think about it?

Perhaps Mr McGuire should have allowed for a second word, too. Derivatives. Or perhaps he should have substituted 'derivatives' and omitted 'plastics' entirely. By the end of 1967, work was already well advanced in the development of a preference-free option pricing formula, based on the work of Sprenkle, Boness, Samuelson, and Thorp. Merton, Black and Scholes were eventually to become household names (well, almost) after their revolutionary results were published in 1973. It is widely acknowledged that financial markets were forever changed by wide adoption of derivatives as a fundamental tool of risk allocation, as recognition of the importance of arbitrage-free option pricing spread; see MacKenzie's (2003) history.

Alas, every boon is also a bane. In the crash of 1987, portfolio insurance strategies, based on option replication arguments, were given much of the blame. Following the various studies of the crash, regulatory authorities prescribed additional regulation over the function of equity markets, in the form of 'circuit breakers', to mitigate future risk. Against this backdrop the Basel Committee on Banking Supervision, in 1988, ushered in a new era of international cooperation in international financial regulation. And the Committee's attention immediately turned to derivatives, motivated by a concern for the soundness and stability of the international financial system. Also in 1988, the Chicago Mercantile Exchange adopted a system for setting daily margin requirements, the Standard Portfolio Analysis System (SPAN®), based on a type of stress test discussed below, scenario analysis. SPAN has since been adopted by many derivatives exchanges.

In the 1990s a series of corporate financial collapses occurred which were associated (in one way or another) with derivatives usage. Some of the names are familiar, such as Orange County (see Jorion, 1995), Procter and Gamble, Gibson Greetings (see Overdahl and Schachter, 1995) and

Barings. Others are less familiar, such as Metallgesellschaft (see Culp and Miller, 1999), Sumitomo (Gilbert, 1996), Daiwa Bank, and SK Securities (see Gay *et al.*, 1999).

As these events unfolded, regulatory authorities sought ways to enhance internal risk controls. In the series of reports and rule makings that resulted, recommendations for best practice frequently included a reference to the importance of the role of stress testing for identifying otherwise hidden risks. It is these recommendations that have pushed stress testing to a relatively prominent position among risk management tools.

Recommendation 6 of the G-30 report (Global Derivatives Study Group, 1993) states:

‘Dealers should regularly perform simulations to determine how their portfolios would perform under stress conditions. Simulations of improbable market environments are important in risk analysis because many assumptions that are valid for normal markets may no longer hold true in abnormal markets. These simulations should reflect both historical events and future possibilities. Stress scenarios should include not only abnormally large market swings but also periods of prolonged inactivity. The tests should consider the effect of price changes on the mid-market value of the portfolio, as well as changes in the assumptions about the adjustments to mid-market (such as the impact that decreased liquidity would have on close-out costs). Dealers should evaluate the results of stress tests and develop contingency plans accordingly.’

The US Comptroller of the Currency in Banking Circular 277 (Comptroller of the Currency, 1993) states:

‘National banks’ ... systems also should facilitate stress testing and enable management to assess the potential impact of various changes in market factors on earnings and capital. The bank should evaluate risk exposures under various scenarios that represent a broad range of potential market movements and corresponding price behaviors and that consider historical and recent market trends.’

The Basel Committee on Banking Supervision (1994) states:

‘Analysing stress situations, including combinations of market events that could affect the banking organisation, is also an important aspect of risk measurement. Sound risk measurement practices include identifying possible events or changes in market behaviour that could have unfavourable effects on the institution and assessing the ability of the institution to withstand them. These analyses should consider not only the likelihood of adverse events, reflecting their probability, but also ‘worst-case’ scenarios. Ideally, such worst-case analysis should be conducted on an institution-wide basis by taking into account the effect of unusual changes in prices or

volatilities, market illiquidity or the default of a large counterparty across both the derivatives and cash trading portfolios and the loan and funding portfolios.’

The Derivatives Policy Group (1995) included stress testing among the necessary risk measurement tools of derivatives dealers. Specifically, they state, ‘Mechanisms should be in place to measure market risk consistent with established risk measurement guidelines. These procedures should ... provide the information necessary to conduct “stress testing”.’

Concurrent with the implementation of the use of VaR models for the calculation of regulatory market risk capital as permitted by the EU Capital Adequacy Directive, banks were required to ‘conduct a routine and rigorous programme of stress testing’ (Securities and Futures Authority, 1995).

The Basel Committee on Banking Supervision (1996) eventually made stress testing a prerequisite for banks to be eligible for the internal models approach to market risk capital. The requirement had previously been laid out by the Committee in an earlier release (1995). More specifically, the Basel Committee on Banking Supervision (1995) states:

‘Banks that use the internal models approach for meeting market risk capital requirements must have in place a rigorous and comprehensive stress testing program. Stress testing to identify events or influences that could greatly impact banks is a key component of a bank’s assessment of its capital position. Banks’ stress scenarios need to cover a range of factors that can create extraordinary losses or gains in trading portfolios, or make the control of risk in those portfolios very difficult. These factors include low-probability events in all major types of risks, including the various components of market, credit, and operational risks. Stress scenarios need to shed light on the impact of such events on positions that display both linear and non-linear price characteristics (i.e. options and instruments that have options-like characteristics). Banks’ stress tests should be both of a quantitative and qualitative nature, incorporating both market risk and liquidity aspects of market disturbances. Quantitative criteria should identify plausible stress scenarios to which banks could be exposed. Qualitative criteria should emphasize that two major goals of stress testing are to evaluate the capacity of the bank’s capital to absorb potential large losses and to identify steps the bank can take to reduce its risk and conserve capital. This assessment is integral to setting and evaluating the bank’s management strategy and the results of stress testing should be routinely communicated to senior management and, periodically, to the bank’s board of directors. Banks should combine the use of supervisory stress scenarios with stress tests developed by banks themselves to reflect their specific risk characteristics. Specifically, supervisory authorities may ask banks to provide information on stress testing in three broad areas, which are discussed in turn below.

‘(a) Supervisory scenarios requiring no simulations by the bank

‘Banks should have information on the largest losses experienced during the reporting period available for supervisory review. This loss information could be compared to the level of capital that results from a bank’s internal measurement system. For example, it could provide supervisory authorities with a picture of how many days of peak day losses would have been covered by a given value-at-risk estimate.

‘(b) Scenarios requiring a simulation by the bank

‘Banks should subject their portfolios to a series of simulated stress scenarios and provide supervisory authorities with the results. These scenarios could include testing the current portfolio against past periods of significant disturbance, for example, the 1987 equity crash, the ERM crises of 1992 and 1993 or the fall in bond markets in the first quarter of 1994, incorporating both the large price movements and the sharp reduction in liquidity associated with these events. A second type of scenario would evaluate the sensitivity of the bank’s market risk exposure to changes in the assumptions about volatilities and correlations. Applying this test would require an evaluation of the historical range of variation for volatilities and correlations and evaluation of the bank’s current positions against the extreme values of the historical range. Due consideration should be given to the sharp variation that at times has occurred in a matter of days in periods of significant market disturbance. The 1987 equity crash, the suspension of the ERM, or the fall in bond markets in the first quarter of 1994, for example, all involved correlations within risk factors approaching the extreme values of 1 or –1 for several days at the height of the disturbance.

‘(c) Scenarios developed by the bank itself to capture the specific characteristics of its portfolio

‘In addition to the scenarios prescribed by supervisory authorities under (a) and (b) above, a bank should also develop its own stress tests which it identifies as most adverse based on the characteristics of its portfolio (e.g. problems in a key region of the world combined with a sharp move in oil prices). Banks should provide supervisory authorities with a description of the methodology used to identify and carry out the scenarios, as well as with a description of the results derived from these scenarios.

‘The results should be reviewed periodically by senior management and should be reflected in the policies and limits set by management and the board of directors. Moreover, if the testing reveals particular vulnerability to a given set of circumstances, the national authorities would expect the bank to take prompt steps to manage those risks appropriately (e.g. by hedging against that outcome or reducing the size of its exposures).’

Another round of calls for stress testing, including recommendations for stress tests of counterparty credit exposures, followed the LTCM and liquidity crises of 1998. See for example, the report of the President’s Working Group on Financial Markets (1999). The Basel Committee

provided further impetus to the application of stress testing to credit in 2002 as part of the new Basel Capital Accord ('Basel II'), stating (Basel Committee on Banking Supervision (2002): 'Banks adopting an [internal ratings based] approach to credit risk will be required to perform a meaningfully conservative credit risk stress test of their own design with the aim of estimating the extent to which their IRB capital requirements could increase during such a stress scenario. Banks and supervisors will use the results of such stress tests as a means of ensuring that banks hold a sufficient capital buffer under Pillar Two of the new Accord.' Credit stress testing is discussed in the Credit Risk section of the *PRM Handbook*.

III.A.4.3 Conceptual Context

As noted in the Introduction, the apparent presumption in all this emphasis on stress testing is that using stress tests will lead to better decisions with respect to risk taking, either as an integral component of the decision maker's objective function, or as a tool measuring the distance from some goal. What is not at all clear, however, is what are the mechanisms at work here.² To put this in the form of a question, if we develop a system of stress testing, what are we then actually supposed to do with the stress-test results anyway?

Berkowitz (1999) takes up this question.³ It is presumed that stress tests are needed because there is something lacking in the VaR derived from the firm's risk model. By 'risk model' Berkowitz means the computational process by which VaR is ultimately derived. Stress tests, if they address this lack, could be thought of as a way of enhancing the assessment of portfolio risk, specifically by providing a way of correcting VaR. To achieve this correction, stress tests should be incorporated into the risk model. He argues that there is no reason to support the modelling of stress tests as a thing separate and apart from the normal process used to evaluate risk, for then there is no internally consistent way to put stress results to use. (Of course, one could still view and use stress-test results in any arbitrary way.)

It is easiest to see how this approach to employing stress tests would work by thinking about the historical simulation approach to VaR estimation (see Chapter III.A.2). In (the basic application of) historical simulation, each day of history is assigned an equal probability in the forecast portfolio return distribution. A set of stress scenarios may then be added to the history, each scenario weighing equally with each day of history. The resulting 'corrected' simulation is then used in evaluating the usual desired percentile for the calculation of VaR, now the 'corrected'

² See Shaw (1997) for a critique of stress tests.

³ See also Zangari (1997a, 1997b) and Cherubini and Della Longa (1999), who apply the Black-Litterman/Bayesian approach to incorporate stress scenarios into the VaR framework.

VaR. The risk model employed by Algorithmics, the enterprise-wide risk management software company, follows a similar approach.

This is a powerful and attractive argument for placing stress tests on a sound footing for use in risk management. The key prerequisite for accepting this approach is a willingness to assign to each stress-test scenario a subjective probability (it is not necessary to assume equal probabilities). While putting stress testing on a solid foundation, this approach is not a panacea for the risk manager. Even were this approach to be adopted, the risk manager would still be left to deal with the anxiety and doubt over including inappropriate scenarios and excluding overlooked scenarios. Interestingly, too, the Basel Committee's risk-based capital rules may create a disincentive for banks to adopt this integrated approach in favour of a fuzzier application of stress testing. Integrating stress tests with the basic risk model will result in an increase in measured VaR, which will result in an increase in regulatory capital. To the extent that capital requirements are a binding constraint on the bank (or nearly so), the bank will incur some real incremental economic costs using this approach.

Is that it, then? Is that the only application of stress testing that is sensible? Well, no, even if it is the most theoretically comforting (i.e., internally consistent) application. As a risk yardstick, stress tests also provide information to decision makers about risk taking in relation to risk appetite. For this reason some institutions set stress-test limits at the enterprise level. Others use stress-test results to measure capital usage or assess capital charges. Less formally, many institutions rely on stress-tests results as a means for the decision maker to perform an intuitive check on his or her comfort with the set of risks in the portfolio. In this sense, stress tests can provide a 'reality check' or a way of assessing model risk (e.g., the sensitivity of valuation to parameter estimates or inputs) where complex models are used in risk assessment. This use of stress testing is long-established practice in the credit world. Regulators also view stress tests as a way to assess their own comfort level with the risks being run individually by institutions for which they are responsible, and more recently also to check their comfort level with the systemic risks implied by the collective positions of the same institutions.

III.A.4.4 Stress Testing in Practice

Most of the information available on current practice was obtained from a survey of 43 large financial institutions conducted by a task force of G-10 central banks established by the Committee on the Global Financial System (2001). The results were also summarised by Fender *et al.* (2001) and Fender and Gibson (2001a, 2001b). The survey asked risk managers to list the most important stress scenarios used firm-wide.

The task force identified nine stress-scenario themes:

- four themes related to asset class (e.g., stress tests on commodity indices);
- four themes related to geographic region (e.g., stress tests on exposures to emerging markets).

The most common scenarios were the 1987 stock market crash, widening of credit spreads, and various versions of a hypothetical stock market crash. Among interest-rate themed scenarios, the bond market crash of 1994 was the most common. In the emerging markets theme, an Asian crisis was the most common. US dollar weakness/strength scenarios were the most common among the remaining geographic themed scenarios. Commodity themed scenarios focused most often on a potential Middle East crisis. Stress tests, other than those centred on foreign exchange, tended to reflect the predominance of long exposures in the respondents' portfolios (e.g., institutions conducted more spread widening than narrowing scenarios).

Banks are not terribly keen to disclose much about their stress-testing approach. Below are produced extracts from the 2003 annual reports of the three largest US banks (by assets).

Citigroup's report states: 'Stress testing is performed on trading portfolios on a regular basis ... on individual trading portfolios, as well as on aggregations of portfolios and businesses, as appropriate. It is the responsibility of independent market risk management, in conjunction with the businesses, to develop stress scenarios, review the output of periodic stress testing exercises, and utilize the information to make judgments as to the ongoing appropriateness of exposure levels and limits.'

Bank of America states: '[S]tress scenarios are run regularly against the trading portfolio to verify that, even under extreme market moves, we will preserve our capital; to determine the effects of significant historical events; and to determine the effects of specific, extreme hypothetical, but plausible events. The results of the stress scenarios are calculated daily and reported to senior management ... ?

JP Morgan Chase, which typically provides the most comprehensive disclosures, states: 'Stress testing ... is used for monitoring limits, cross business risk measurement and economic capital allocation. ... The Firm conducts economic value stress tests for both its trading and its non-trading activities, using the same scenarios for both. The Firm stress tests its portfolios at least once a month using multiple scenarios. Several macroeconomic event-related scenarios are evaluated across the Firm, with shocks to roughly 10,000 market prices specified for each scenario. Additional scenarios focus on the risks predominant in individual business segments

and include scenarios that focus on the potential for adverse moves in complex portfolios. Scenarios are continually reviewed and updated to reflect changes in the Firm’s risk profile and economic events. Stress-test results, trends and explanations are provided each month to the Firm’s senior management and to the lines of business, to help them better measure and manage risks and to understand event risk-sensitive positions. The Firm’s stress-test methodology assumes that, during an actual stress event, no management action would be taken to change the risk profile of portfolios. This assumption captures the decreased liquidity that often occurs with abnormal markets and results, in the Firm’s view, in a conservative stress-test result. ... [S]tress-test losses are calculated at varying dates each month. ... The following table represents the worst-case potential economic value stress-test loss (pre-tax) in the Firm’s trading portfolio as predicted by stress-test scenarios:

Trading Economic-Value Stress-Test Loss Results – Pre-Tax

as of or for the year ended December 31

Loss in USD \$m

2003				2002(A)			
AVE.	MIN.	MAX.	DEC. 4	AVE.	MIN.	MAX.	DEC. 5
508	255	888	436	405	103	715	219

(A) Amounts have been revised to reflect the reclassification of certain mortgage banking positions from the trading portfolio to the non-trading portfolio.

“The potential stress-test loss as of December 4, 2003, is the result of the “Equity Market Collapse” stress scenario, which is broadly modeled on the events of October 1987. Under this scenario, global equity markets suffer a sharp reversal after a long sustained rally; equity prices decline globally; volatilities for equities, interest rates and credit products increase dramatically for short maturities and less so for longer maturities; sovereign bond yields decline moderately; and swap spreads and credit spreads widen moderately.”

III.A.4.5 Approaches to Stress Testing: An Overview

The definition and practice of stress testing encompass several different techniques (see Table III.A.4.1 for a high level overview). The choice of which test to employ in a particular institution and situation is driven by several factors. Regulatory requirements are important, of course. Also important is the cost, in time and resources, needed to generate stress-test results. The choice also reflects the specific needs of the users. These needs depend on the complexity of the

portfolio, the frequency with which it is traded, the liquidity of the instruments in the portfolio, the volatility of the markets in which the instruments are traded, and the strategies employed.⁴

Table III.A.4.1: Typology of stress tests

Approach	Description	Pros	Cons
Historical scenarios	Replay crisis event	It actually happened that way	Proxy shocks may be numerous No probabilistic interpretation No guarantee of ‘worst case’
Hypothetical scenarios	1. Covariance matrix 2. Create event 3. Sensitivity analysis	1. Relatively easy 2. Very flexible 3. Can be detailed	1. Empirical support mixed 2. No guarantee of ‘worst case’ 3. Limited risk information
Algorithmic	1. Factor push 2. Maximum loss	1. Minimal qualitative elements 2. Identifies ‘worst case’ in feasible set (maybe)	1. No guarantee of ‘worst case’ 1. Ignores correlations 2. Assumes data from normal periods are relevant 2. Computationally intensive

Most of the regulatory attention has focused on stress testing at the portfolio level. For the regulators it is the aggregated impact of stressed market environments that poses risks that interest them. For some time international organisations have pursued the idea of aggregating the results of standardised stress scenarios for individual financial firms into financial sector stress tests to attempt to estimate the economy-wide impact of crisis events (e.g., the Committee on the Global Financial System of the Bank for International Settlements, and the Financial Sector Assessment Program of the IMF and World Bank). Stress testing of a sort is widely practised at the individual trading desk, trader and position levels, too. To the extent that trading desks already perform their own position-by-position stress testing, we have to ask why we cannot merely aggregate those results for the purpose of examining exposure at the portfolio level. Desk-level stress tests are (as they should be) highly focused on the specific risks being run at the desk. Perhaps much of the information generated is not relevant at the aggregate portfolio level. Equally important, risks that are deemed to be of ‘second order’ at the trading desk, and hence are subject to little or no stress testing, may be important (at the margin at least)

⁴ The discussion that follows pertains most directly to stress testing for traded instruments. Stress testing either the banking book or a non-financial firm’s exposures presents additional challenges and needs. Approaches specific to those needs have been developed (e.g., net interest income stress tests and economic value added stress tests), but they are outside the scope of this discussion.

when viewed in the context of risks being taken at other desks. Also, desk-level tests are used for evaluating and actively managing risk position by position (or at least strategy by strategy), possibly in near real time. Portfolio stress testing serves a more strategic purpose in the identification and control of event-related risk, and the most useful metrics for the one purpose need not also serve best for the other. Desk-level stress tests tend to be more of the factor-push variety, whereas portfolio stress tests tend to employ the scenario approach.

Irrespective of whether the stress-testing programme is intended for desk level or portfolio level risk management, useful stress tests require full revaluation for all positions with nonlinear or discontinuous payouts.⁵ Perhaps it is foolish to state the obvious. However, full revaluation can entail significant computational and time costs, and it is tempting to trade off accuracy for speed. But, as the *raison d'être* of stress testing is to explore portfolio losses in crises where nonlinearities and discontinuities may expose hidden risks, approximate revaluation approaches should be strictly limited. It may be necessary to develop alternative revaluation strategies to overcome technology constraints that limit the employment of full revaluation for other risk management purposes. The only exception to this requirement is for some desk-level stress testing. Here stress testing of sensitivity-based risk exposure information (deltas, vegas, etc.) is useful for tactical portfolio management decisions as well as being perhaps the only practical way of delivering stress results on demand or in real time.

III.A.4.6 Historical Scenarios

Historical scenarios, as a variety of stress testing, seek to quantify potential losses based on re-enacting a particular historical market event of significance. Scenario shocks that determine the impact on portfolio valuation are taken from observed historical events in the financial markets. This is in contrast to stress tests where shocks are based, for example, on specified changes in the covariance matrix of asset returns, or on shifting prices or rates by an arbitrary number of standard deviations. We are all empiricists when we say, 'it is reasonable, because it actually happened'. Historical scenario stress testing is required by the Basel Committee, prescribing that 'scenarios could⁶ include testing the current portfolio against past periods of significant disturbance, for example, the 1987 equity crash, the ERM crises of 1992 and 1993 or the fall in bond markets in the first quarter of 1994, incorporating both the large price movements and the sharp reduction in liquidity associated with these events' (Basel Committee on Banking Supervision, 1996).

⁵ Discontinuities can present a challenge for stress tests of all types, except perhaps in principle 'maximum loss'. It may be prudent to construct stress tests with the points of discontinuity specifically in mind.

⁶ With regulators, to say 'could' is to say 'do this unless you have some compelling reason not to'.

It sounds so easy. Designate a period in history with a suitable crisis environment and find out what would happen to the current portfolio if that crisis were replicated today. However, as is the case with so many things, it really is necessary to ‘sweat the details’. I will consider the elements of historical stress testing in turn, namely, choice of a historical period and specifying shocks.

III.A.4.6.1 Choosing Event Periods

The first question in choosing a historical period for stress testing is which periods to choose. A historical event may be defined in one of two ways. In the first, the event is defined relative to a well-known crisis period, such as the Asian crisis of 1997. In the second, the event is defined by examining the historical record of moves in market risk factors relative to some user-defined threshold level of shocks. The second approach will no doubt also turn up events that correspond to most well-known crises, but may identify other event periods as well, depending on the particular risk factors whose histories are being scanned for large movements. The former approach is more prevalent.

Some potential scenarios were mentioned in previous sections. Common candidates for historical stress tests include the following: the US stock market crash of 1987, the European exchange-rate mechanism crisis of 1992, the US bond market sell-off of 1994, the Mexican peso crisis of 1994, the so-called Asian crisis of 1997, the Russian default of 1998, and the LTCM and liquidity crises of 1998. The attacks of 11 September 2001 also constitute a candidate.⁷ Davis (2003) creates a typology of financial crises. Systematic characterisation of market crises may be useful both for ensuring that the set of events employed in a stress-testing programme contains a reasonable portion of the spectrum of possible crises, and for guiding the construction of hypothetical scenarios (discussed later).

The next question is how many events should be part of the stress-testing programme. No matter how many historical periods are selected, it is not possible to guarantee that the prospective worst-case scenario is covered. It can be hoped, however, that a judicious selection of scenarios will provide indicative information about areas of vulnerability, especially in identifying risks that may not be obvious from other risk measurements, such as VaR. Ideally, the risk manager will vary the number and variety of historical scenarios evaluated through time depending on both the changing composition of the portfolio and the changing economic environment.

⁷ See, for example, Malz and Mina (2001), Hickman and Jameson (2001) and Jorion (2002).

The first thing one realises when looking at a historical event candidate for a stress scenario, say the liquidity crisis of 1998, is that the start and end dates of the event are not always obvious. For any particular market rate, the period of interest is likely to be unambiguous. However, the more complex and varied the instruments in the portfolio to be stressed, the more difficult is the problem of identifying the start and end date for the stress event. Two approaches are possible.

- Define the event interval, making sure that the interval selected encompasses all (or essentially all) of the significant moves in individual market rates. Then use as the shock magnitude for each risk factor the greatest change in that factor (e.g., peak-to-trough) found within the interval, regardless of the start or end date. The advantage of this is that the scenario will entail the largest possible moves in each risk factor. The disadvantage is that the shocks, when taken together, may make no economic sense.
- Define the event interval such that it comes as close as is possible to capturing exactly the greatest moves in the factors of most interest. That is, the peaks (or troughs) occur at the start date and the troughs (or peaks) occur at the end date. Then use as the shock magnitude for each risk factor the change in that factor from the start date to the end date. The ideal event window cannot be achieved in practice. That is the disadvantage of this. The advantage is that the scenario has the potential to be economically meaningful.

The second approach is preferred, as a key element in establishing the plausibility of a scenario is that the shocks, taken together, must be sensible.

Historical scenarios rarely play out within a single trading day. Given international market linkages observed through contagion and feedback effects, even an event sharply focused in time is likely to engender after-shocks that continue for a few days. More commonly, an event will develop over a period of a few days or even weeks, as in the liquidity crisis of 1998. As a result the specification of historical events raises questions about how the passage of time is affecting the test results. Two areas of concern come to mind: trading or hedging out risks, and modelling the effect of time passing on expiration, maturity and ‘carry’ (e.g., for fixed income or options positions).

- No matter how illiquid a market, there is a price at which a portfolio manager can trade out of a position. Since those costs may be prohibitive, it is common to assume in historical stress scenarios that positions cannot be traded or hedged – no matter how long the interval of calendar time spanned by the scenario. It is common to argue (at least, I have argued) that this assumption, while extreme, approximates a worst-case scenario in which illiquidity is extreme. Still, it stretches the plausibility requirement to apply this assumption to a very extended stress period (as traders will readily argue).

- Another question that arises with historical scenarios is how to model the impact of the passage of time on instruments in the portfolio whose values depend on time. Instruments that are affected include futures and forwards, bonds and options. It is possible to argue that the scenario telescopes the historical record into a single trading day, thereby making it unnecessary to deal with the passage of time, and this is the prevalent approach. However, by assuming that time is telescoped, the effects of illiquidity on the portfolio are muddled somewhat, and the plausibility of such assumed one-day moves in rates can be questioned. If explicitly allowing for the passage of time, the expiration of options and the cash flows from payouts, if any, need to be incorporated into the scenario. Similarly, bond coupons and repo payments should be considered. Allowance must be made, too, for rolling of forwards and futures.

III.A.4.6.2 Specifying Shock Factors

A fundamental element in specifying shocks for a historical scenario is the choice of relative (or proportional) versus absolute (or additive) shocks. Consider the following example. Suppose that the GBP/USD exchange rate is currently 1.8. Also assume that during the period of historical interest the rate moved from 1.5 at the beginning of the event to 1.75 at the conclusion of the event. The absolute change in the exchange rate during the event was 0.25, and the relative change was 16.7% (appreciation of the GBP). A decision must be made whether to calculate the shocked exchange rate as $1.8 + 0.25 = 2.05$ or $1.8 \times 1.167 = 2.1$. This issue arises in VaR modelling as well, but because the changes in market risk factors are generally larger in stress testing, the effects can be more dramatic. If the levels of market rates are similar between the beginning of the historical event and the current stress-test 'as of' date, then the choice is less important. However, this will not generally be the case.

Relative shocks are generally preferred for a couple of reasons. Firstly, when applying a relative shock one will not inadvertently cause a rate to change sign. Secondly, a relative shock (generally) corresponds directly to the rate of return on a portfolio, which is (generally) thought to be the parameter of interest in an individual's utility function. Nevertheless, relative shocks are not always appropriate. Some market spreads can be either positive or negative. To maintain that property in a stress scenario, absolute shocks need to be applied. Applying relative shocks to interest rates can be a problem as well. For example, a relative interest-rate shock that is derived from a historical period when interest rates are very low may imply unrealistic moves in rates when they are at higher levels. As a general rule, then, most shocks should be relative. However, shocks to interest rates generally should be absolute. Shocks to volatility should generally be relative, in part because volatilities cannot be negative. Exceptions to these rules should be made on a rate-by-rate basis where it is appropriate to do so.

An issue arises when the portfolio's fixed-income instruments are priced from both zero curves and par curves, perhaps as a result of the way different trading desks prefer to view their risks, or when different portfolios are marked using different back-office systems. In this case, consistency between the historical shocks for par curves and zero curves must be imposed.

Applying historical interest-rate shocks to current yield curves requires special attention as well. Years of research (and probably quite a few doctoral theses) have shown that typically three factors can explain about 95% of the movements in yield curves, commonly interpreted as level, slope and curvature. Ideally, it would be the shocks arising from those three key yield curve factors that are used in scenario construction, perhaps through principal components analysis (PCA) as described in Frye (1997). If historical (absolute) changes at various points on the yield curve are applied point-by-point to the current yield curve, the problem that can arise is that the 'shocked' yield curve can take on some very implausible shapes. It is a good idea, at the very least, to monitor the 'look' of shocked curves as part of the stress-test process.

PCA is useful for generating realistic 'shocked' yield curves in a tractable manner. Firstly, dimensions are reduced so that only the three key risk factors need to be shocked. Secondly, these factors are orthogonal, so they can be shocked independently and the shocked result will be a realistic curve. This technique is equally important when the scenario specifies shocks to any term structure, such as the term structure of volatility, a term structure of commodity futures of different maturities and a term structure of foreign exchange rates. If PCA is not applied it may be necessary to modify historical shocks that give rise to implausible curve shapes, perhaps by applying the Cholesky matrix (derived from the historical covariance matrix) to the independent shocks, thus making them correlated (see Section II.D.4).

III.A.4.6.3 Missing Shock Factors

In many instances it is not possible to refer to the historical record for a particular instrument when specifying stress shocks. The more distant in the past is the historical scenario, the more likely this is to be a problem. In some cases instruments in the current portfolio simply were not traded during the historical period. For example, default swaps were not traded at the time of the Mexican peso crisis. In other cases, even if instruments were traded, there is no reliable source for historical data; either the data are bad or there are no data. It is necessary to have a policy for dealing with 'missing' shock factors.

A simple rule to follow is not to leave any current positions without an assigned historical shock unless a zero shock is the best guess at what that shock would have been. If positions in particular instruments are deemed to be immaterial at the time of the specification of the

scenario, the situation may subsequently change, and the specification of the scenario would then need to be revisited. If stress scenarios become part of the risk limit structure or capital allocation or performance evaluation processes, leaving positions without shocks not only results in an inferior decision-making process but also creates incentives for strategic behaviour that may not be desirable from the perspective of risk appetite.

Assumptions about correlations play a big role in following the leave-no-position-unshocked rule, as the best guess is usually obtained from examining historical shocks of instruments thought to be highly correlated with the position in need of a shock assignment. There are two basic approaches to ensuring this: employing proxies or using interpolation.

- In the case of proxies, a shock is assigned from another instrument. Equities are a good example. With equities, mergers, spin-offs, and changes in business focus may make this existing historical record irrelevant, at least if plausibility is to be maintained. In this case it may be prudent to at least ensure that changes in a company's industry classification are noted. Then a policy decision can be made whether to ignore the history and instead proxy shocks for such equities to a historical industry-specific shock factor. When assigning proxies, it is useful to employ more rather than fewer proxies (equity sector proxies, rather than just a single market proxy), in order to obtain a better-articulated result. More sophisticated data filling approaches than discussed here are possible as well, of course.
- Interpolation (or extrapolation) may be appropriate in the case of fixed-income instruments. For example, swap and forward foreign exchange markets tend to both become more liquid and extend over time as comfort increases in assessing the longer-term risks. This filling in and out of the term structure is especially noticeable in emerging markets. In part because of the correlation between instruments of different tenors and in part because of the structure of implied forward rates, it is reasonable to interpolate rate shocks (usually linearly) from available historical shocks at adjacent points in the term structure.

It should be clear that even when using historical scenarios for stress testing, scenario creation is not a once- only event. Not only should new scenarios be constantly under consideration for development, but even existing scenarios need to be constantly re-evaluated and sometimes tweaked to maintain their usefulness. This can be tedious and unexciting, so it is a good idea to establish a policy to formally review stress scenarios periodically to assist in establishing a good discipline.

III.A.4.7 Hypothetical Scenarios

History does not conveniently present the risk manager with a template for every plausible future market crisis (though the sample size of crises does keep increasing with time). For this reason, it may be desirable to create a hypothetical economic scenario as a stress test. Ideally, a hypothetical scenario is based on a structural model of the global financial markets (perhaps with a ‘real’ or physical goods and services component, too), in which the specification of a parsimonious set of market shocks provided as inputs to the model will result in a complete specification of responses in all markets. Well, in most cases that is not going to happen.⁸ Still, it is good to keep that ideal in mind when constructing an economic scenario, because it is very easy to make a bad scenario by ignoring cause, effect and co-determination in economic relationships.

III.A.4.7.1 Modifying the Covariance Matrix

Some argue that the key feature of a stress event is embedded in the behaviour of asset correlations. The intuition is strong. In a crisis, investors may make fewer distinctions among assets and issue blanket buy or sell orders in a flight to safety that tends to drive whole classes of assets in the same direction. Or interconnections between market participants may manifest in a crisis where the actions of one agent create the need for other agents to take similar actions. Or interconnections between markets may manifest when an agent, faced with a liquidity crisis in one market, attempts to liquidate positions in other markets, precipitating a liquidity crisis throughout the system. A stress test can be constructed from a modified covariance matrix in several ways. For example, if it is assumed that asset returns are jointly normally distributed, then the stress portfolio valuation can be obtained by computing the monetary value of a one standard deviation change in the portfolio value (using the modified covariance matrix) and scaling up the result by the desired number of standard deviations (a common multiplier is 4).

As with any intuition, it is prudent to test it against the available data. Boyer *et al.* (1999) show that careless data mining can lead one to conclude incorrectly that correlations are different in stressful markets. However, they acknowledge that there is evidence that correlations do change. Taking Boyer *et al.*'s points into account, Kim and Finger (2000) conduct an empirical study from which they conclude that that data provide considerable support for the existence of a separate stressful market environment with distinct asset correlations. In particular, they propose that observed asset returns are generated by a mixture of normal return-generating processes, one for an ordinary market environment and one for a stressful market environment. Nevertheless,

⁸ However, that is exactly the approach taken by UK regulatory authorities who employed a macroeconomic model in their experimentation with macroeconomic stress tests. Hoggarth and Whitley (2003) present a very interesting discussion of the issues involved.

Loretan and English (2001) argue that the empirical evidence might not support correlation breakdown, but rather be the residual of time-varying volatility. Similarly, Forbes and Rigobon (2002) argue that, taking into account the apparent relation between correlation and volatility, they are unable to find evidence of changes in correlations during the 1997 Asian crisis, the 1994 Mexican peso crisis, or the 1987 US stock market crash. Then again, Dungey and Zhumabekova (2001) and Corsetti *et al.* (2002) say that the results in Forbes and Rigobon (2002) may be overstated, first because the number of crisis periods in the sample is small, and second because their econometric specification is too restrictive.

In sum, despite the intuition, the empirical evidence is not uniformly supportive of the notion that correlations increase in crisis situations. Still, the force of intuition is strong, and it is common to construct stress scenarios with increases in correlation. Note, however, that increased correlation does not in itself guarantee to stress a portfolio. It is simple to construct thought experiments in which the VaR of a portfolio will decline with increases in correlation.

It is not advisable to change correlations by arbitrarily setting selected correlations to 0, 1 or -1 , because implausible stressed portfolio returns can result (specifically, covariance matrices that are not positive semi-definite, meaning that some portfolios could have negative variance!). As a result, any stress scenario that involves causing correlations to differ from the relationships embedded in the historical data that were used to estimate them must follow certain rules.

When changing the correlation between two risk factors, it is important to understand that the correlations can only change in reality if the underlying returns on the two factors change relative to each other (possibly in such a way that the average return and the variance of the two do not change). If those underlying returns change, then by implication every correlation between each of those two risk factors and the remaining risk factors may change, too. Thinking in terms of the correlation matrix, if we want to change the value of one correlation, which means changing the value in two cells of the matrix (from the symmetry property), then all the correlations in the rows and columns intersecting those two cells can change as well. Generally, many possible pairs of altered return vectors will yield any desired correlation (given average returns and variances). By specifying the method to (explicitly or implicitly) adjust the return vectors, it will be possible to determine the corresponding induced changes to the other affected correlations that are necessary to maintain consistency.

Finger (1997) proposes modifying the return vectors of the risk factors whose correlations are to be modified such that the return on each factor on any day, t , is a linear combination of the historically observed return of that factor on day t and the average of the returns on the affected

market factors on day t . The transformed return vectors will need to be rescaled if the original individual variances are to be unchanged. Note that shocking individual asset variances can be incorporated into this method as well. Further, an adjustment may also be made to ensure the mean returns are unchanged, but this is not necessary if the mean returns are ignored (i.e., assumed to equal zero) for purposes of the risk calculations.

Table III.A.4.2 contains statistics derived from the logarithms of the daily changes in closing prices (in New York) for IBM, GE and MSFT for one year ending on 20 April 2004. The upper triangular (red) entries are correlations;⁹ the remaining entries are variances and covariances (all annualised from the daily data).

Table III.A.4.2: Estimated correlations and annualised variances and covariances

	IBM	GE	MSFT
IBM	0.038	0.411	0.558
GE	0.016	0.041	0.474
MSFT	0.025	0.022	0.054

In this example, the correlation between IBM and MSFT will be increased to 0.850 (from 0.558) using Finger’s methodology. Table III.A.4.3 illustrates the operations that are performed on the IBM and MSFT return vectors.

Table III.A.4.3: Modification of returns on a representative date

Date	Return	Modified Return	Normalised Return
4 June 2003	$R(\text{IBM}) = 0.005117$	$R(\text{IBM}_{\text{mod}}) = \theta \times \frac{0.005117 - 0.0004}{2} + (1 - \theta) \times 0.005117$	$R(\text{IBM}_{\text{mod}}) \times \frac{\sigma_{\text{IBM}}}{\sigma_{\text{IBM}_{\text{mod}}}}$
	$R(\text{MSFT}) = -0.0004$	$\text{MSFT}_{\text{mod}} = \theta \times \frac{0.005117 - 0.0004}{2} - (1 - \theta) \times 0.0004$	$R(\text{MSFT}_{\text{mod}}) \times \frac{\sigma_{\text{MSFT}}}{\sigma_{\text{MSFT}_{\text{mod}}}}$

The parameter θ determines the modified correlation between the two return series. For any choice of θ , returns for every date are modified as illustrated for the representative date in the table. If $\theta = 1$, then the two time series will be perfectly correlated. A simple numerical search (e.g., with Solver in MS Excel) can be used to identify the value of θ that results in a correlation equal to the target level of 0.850. For the data used here, the desired correlation is obtained with $\theta = 0.4625$. Table III.A.4.4 shows the final modified correlations and covariances for these three equities.

⁹ For example, $\text{Corr}(\text{IBM}, \text{GE}) = \text{Cov}(\text{IBM}, \text{GE}) / \{\text{Var}(\text{IBM}) \times \text{Var}(\text{GE})\}^{1/2} = 0.016 / \{0.038 \times 0.041\}^{1/2} = 0.411$.

As before, the entries above the diagonal are correlations, and the remaining entries are variances and covariances. As should be expected, given the methodology, the correlations between GE and both IBM and MSFT have been affected by changing the correlation between IBM and MSFT.

Table III.A.4.4: Modified correlations and annualised variances and covariances

	IBM	GE	MSFT
IBM	0.038	0.470	0.850
GE	0.018	0.041	0.498
MSFT	0.038	0.023	0.054

When several instrument pairs are chosen to have their correlations fixed at a level different from the historical correlation, this approach in general requires computing a separate θ for each of those pairs of instruments. In this case, because of the interdependencies among the risk factors, it will be necessary, in general again, to solve simultaneously for the set of θ s that together yield the desired set of correlations.

More general techniques for modifying correlations in a consistent manner have been suggested; see, for example, Kupiec (1998), Rebonato and Jäckel (1999), Higham (2002), or Turkay *et al.* (2003). These approaches focus explicitly on eliminating the negative eigenvalues of the stressed correlation matrix, and seek to identify the consistent matrix that is ‘closest’ to the stressed matrix (according to some metric).

III.A.4.7.2 Specifying Factor Shocks (to ‘create’ an event)

Rather than creating a stress test through modification of the covariance matrix, it is possible to create a hypothetical scenario simply by specifying hypothetical shocks to the market factors. Without an economic model, it is a daunting task to attempt to describe a coherent set of hypothetical shocks encompassing every market risk factor. Thus, even for a hypothetical economic scenario, actual historical behaviour of market prices can provide useful guidance for the specification of plausible shocks.

Another element in specifying shocks is to specify which no-arbitrage relationships are to hold in the scenario. ‘Arbitrage’ is a term that is used very loosely, so much so that a variation has come into use, pure arbitrage. A pure arbitrage is an opportunity for a riskless and certain profit. Pure arbitrage is achieved through the implementation of a self-financing ‘replicating portfolio’ (or exact hedging strategy). Speaking somewhat loosely, instruments that are (in principle) tied by

this type of arbitrage relationship include simple European options and their underlying assets, forwards/futures and the corresponding cash market instrument, and relative exchange rates. The scenario designer must decide which relationships are to be fixed in the scenario. In part, this decision is helped by observing what happened in the historical record. In the case of the relationship between futures and cash equities, it has been observed that the parity relationship did not hold continuously through the 1987 market crash. Nevertheless, if desired, the futures–cash relationship can be enforced in scenario construction simply by defining the futures price to be fairly valued relative to the shocked index level, forgoing the implementation of a separate historical shock for the futures.

Currency traders sometimes will make large bets on the actions of governments that either have pegged or are managing the float of their exchange rate. For this reason, it is important when constructing a hypothetical economic scenario to make a conscious decision about what to do with pegged currencies. Depending on the scenario, it may be appropriate to assume that certain pegs are broken. The shocks that are assumed (including spill-over effects), then, may depend on historical cases when other currency pegs were broken, or on expectations of the future exchange rate implied in non-deliverable currency forwards.

An example of a hypothetical economic scenario is the ‘commodity themed’ Middle East crisis scenario common among banks, as noted in Section III.A.4.2. It may be assumed in such a scenario that the outbreak of war results in a disruption of oil production. Some spike in crude prices and volatility must then be assumed. The impact on related energy products prices is then estimated. However, the impacts do not end there, as such an event is likely to modify investor expectations of future inflationary impacts, lead to possible shortages in certain markets, and bring about pre-emptive central bank responses, etc., all of which will affect asset prices and volatilities. For all of these effects factor shocks must be assumed in a way that creates a plausible and coherent picture of the impacts on various markets. These shocks to market prices and rates are then applied to portfolio positions to evaluate potential exposure to the hypothetical event.

III.A.4.7.3 Systemic Events and Stress-Testing Liquidity

Another hypothetical economic scenario that is of great interest is a systemic liquidity event. The stress tests discussed above do not probe vulnerabilities arising from the interrelationships among institutions. The voluntary withdrawal of a derivatives dealer from the market is a commonly cited event of this type; see, for example, Greenspan (2003) and Jeffery (2003). Since the LTCM and liquidity crises of 1998, regulators and some money centre banks have shown increasing interest in the both immediate and the follow-on effects of such an event. The focus

of this interest is on the mechanisms that tie together institutions and provide channels for contagion and feedback. Borio (2000) writes that, ‘for a proper understanding of liquidity under severe stress, the interaction of basic order imbalances with cash liquidity constraints and counterparty risk needs to be explained. Leverage and risk management play a key role. It also suggests that some factors that may contribute to liquidity in normal times can actually make it more vulnerable under stress.’

Consider the following features of the financial system, all of which are generally regarded as improving stability and efficiency in normal markets. Firstly, collateral requirements for over-the-counter derivatives and performance bonds on exchange-traded derivative transactions reduce the likelihood of default in normal times. However, an organisation that trades these may plan to have liquidity sufficient for ‘normal’ daily mark-to-market contingencies and long-term average liquidity needs, but may still find itself unable to post the collateral as required in a significant market event. Failure to do so requires the counterparty absorb a portion of the loss. This loss may, in turn, force the counterparty to fail to make required payments on its obligations, creating a contagion of defaults.

Secondly, risk management policies, such as risk limits, will moderate risk taking to tolerable levels in normal times. However, where a market event causes a sharp increase in measured risk, traders may choose (or be directed) to exit positions to avoid triggering risk limits. The resulting trades may contribute further volatility, and further knock-on effects. This likelihood is greater if large market participants measure and limit risk taking in similar ways, or risk managers at large firms just respond to market events in similar ways. Then these individual responses will be exacerbated at the level of the financial system as a whole.

Thirdly, position transparency and risk disclosures, as well as advances in information technology, generally promote efficient price discovery and fair market pricing. However, they may also contribute to ‘herd’ behaviour, in which large market players have similar trades, and in which market participants react similarly and at simultaneously in response to an event. This behaviour can exacerbate the initial effects of an event.

Fourthly, a linchpin of efficient functioning of the over-the-counter derivatives market is the ability of dealers to control portfolio risk exposures through the construction of synthetic hedges (replicating the desired set of risk characteristics through another portfolio of instruments). Illiquid sectors of the market rely on the instruments traded in the more liquid sectors to create hedges cheaply. A reduction in liquidity in the liquid sectors can significantly affect the prices in the illiquid sectors. Thus pricing is interrelated across the range of derivatives products. If a

large dealer were to pull back from trading, the resulting reduction in market liquidity would create losses at other institutions, as markets repriced to reflect the reduced liquidity. In some instances, less liquid instruments might become uneconomical to trade, forcing institutions to exit those positions, causing further price impacts in a cascading effect.

The data needs for an institution to construct such a stress test are great. But, as with any stress-testing method, value-added information is still possible even though the reality falls short of the ideal. Firstly, an institution must first incorporate counterparty information in its risk database. Secondly, an institution should be able to identify the impact on required collateral of a proposed set of systemic shocks. This may be especially difficult for institutions with many types of instruments held with a prime broker. Prime brokers will take a portfolio view of their counterparty risk and the algorithm they apply for determining collateral may not be transparent. Thirdly, an institution should estimate the distribution of positions across the financial system. For example, the institution should ask whether dealer A is the only dealer making a market in certain instruments in the institution's portfolio; who are the others and what is the market share of each; and whether the market is one-sided (e.g., the dealer is long in comparison to most counterparties and is primarily relying on hedges to control overall risk) or two-way (the dealer's book has a balance of long and short positions). For the instruments that the institution uses in its own hedging, it should identify the other major institutions that either make markets in those instruments or heavily employ those instruments and estimate market share. Since much of this information is not internal to the institution, some estimation will be necessary. Some potentially useful sources of data for this type of stress test are BIS statistical summaries, reports of derivatives activities of banks published by the US Comptroller of the Currency, and the Commodities Futures Trading Commission commitments of traders' reports.

As is the case with data, specifying shocks can be a significant challenge as well. Some historical guidance is available by reviewing the behaviour of markets and spreads in periods of illiquidity, such as October 1998. It may be possible to use the institution's own trading data to estimate market impact functions in certain instruments. An impact function measures the cost of trading as a function of position size, given other parameters that describe the trading environment (e.g., volume and volatility). The institution's own pricing models may be used to estimate the price impact on positions of a wider bid-ask spread. Ultimately, it will be necessary to rely heavily on intelligent guesstimates of possible shocks.

Supranational organisations that have an interest in the financial stability of the global economy are embracing a similar approach to stress testing, the most prominent bit being the Financial Stability Assessment Program. The umbrella organisation for these efforts is the Financial

Stability Forum (www.fsforum.org), with the actual work being undertaken by the International Monetary Fund (www.imf.org) and the World Bank (www.worldbank.org) in conjunction with local country supervisory authorities. See, for example, Blaschke *et al.* (2001) and International Monetary Fund (2003) for an overview of this effort.

III.A.4.7.4 Sensitivity Analysis

Sometimes it may be desirable to create simple, somewhat artificial portfolio shocks, in a method referred to as sensitivity analysis. In this type of stress test at most a few risk factors are shocked and correlation is typically ignored. These are easy to implement, but they only provide a partial picture, and must be accompanied by a lot of judgement on the part of the risk manager. Examples of sensitivity analysis are a parallel shift of the yield curve, or a 10% drop in equity prices. The Derivatives Policy Group (1995) report contains recommendations for a parallel yield curve shift of 100 basis points, and for curve steepening and flattening of 25 basis points.¹⁰

Since the biggest losses do not always correspond to the largest moves in factors, it is common in scenario analysis to create a ‘ladder’ of shocks, in which price impacts are calculated for intermediate values of the risk factors. Design of these ‘sensitivity ladders’ requires attention to two issues: granularity and range.

The range of shocks should be wide enough to encompass both likely and unlikely (but plausible) moves in the market factors. The recommendations of the Derivatives Policy Group noted above might have represented an adequate range in 1995, but would probably be considered inadequate in 2003–2004, when interest-rate volatility was very high. Similarly, the range selected should take into account the time horizon for the analysis, with the range increasing with the horizon (except, perhaps, for very strongly mean-reverting risk factors). For these reasons, it makes sense to use the volatility (or perhaps, empirical percentiles of the percentage change) of the factor to determine the range.

The granularity of the analysis refers to the distance between the rungs of the ladder, or more exactly, the increment chosen for the change in the risk factor. Granularity should reflect the nature of the portfolio. A portfolio whose valuation function is linear in the risk factor can have

¹⁰ Implementing even these simplistic hypothetical market moves contains some hidden issues that can have a big impact on the results. For example, to implement changes in curve slope, a point on the curve must be chosen around which to rotate the curve, and a second point on the curve must be chosen from which to measure the amount of steepening or flattening. These points should reflect the market environment and the manner in which the curve is traded by market participants. A fair place to start might be to take the two-year point as the point of rotation and the 10-year point as the point to measure the change in basis points.

a larger increment than a highly nonlinear portfolio with some instruments whose payouts might be discontinuous in the risk factor.

III.A.4.7.5 Hybrid Methods

Kupiec (1998) proposes a methodology that is a particular hybrid of covariance matrix manipulation and economic scenarios. His approach can also be applied to the problem of missing historical data in specifying shocks to be used in a historical scenario. In his approach, which he calls ‘stress VaR’, the risk manager ‘can specify partial “what if” scenarios and use the VaR structure to specify the most likely values for the remaining factors in the system’.

Assume that the risk manager wishes to specify the shocks to a subset of the market risk factors. Using the covariance matrix of factor returns and this subset of fixed shocks, it is then possible to compute a conditional mean vector and a conditional covariance matrix for the remaining risk factors. Using the resulting conditional distribution of factor returns (the factors with fixed shocks have means given by the shocks, zero variance, and hence zero correlations), a conditional, ‘stress’ VaR can then be calculated. Kupiec goes on to demonstrate how this approach can be generalised to the case in which selected variances and covariances are given prespecified shocks as well. Assuming that market risk factors are jointly normally distributed makes the approach very tractable.

Using the unconditional covariance matrix as the starting point for this approach is potentially very limiting, if it is true that in crisis periods there is a structural change in the relationships among market risk factors. However, in applying this approach, it is not necessary to create the conditional return distribution from the same covariances as are used in an unstressed VaR. Any covariance matrix may be taken as a starting point, such as the historical covariance matrix modified in the manner of Finger (1997) discussed above. Alternatively, Kim and Finger (2000) employ Kupiec’s method after first estimating a stress-environment covariance matrix. They assume that returns are drawn from a mixture of two normals, one the stress environment and the other a normal environment, calculating the ‘stress VaR’ using the estimated parameters of the stress-environment return distribution. Perhaps more simply, the stress VaR could be calculated using the covariance matrix estimated from a particular crisis period.

III.A.4.8 Algorithmic Approaches to Stress Testing

One of the deficiencies of historical and hypothetical stress scenarios is that the user has only a fuzzy level of confidence that the full extent of the potential badness for the portfolio has been exposed. A more systematic approach might yield a greater level of comfort. The goal is to

create a search algorithm to identify the worst outcome for the portfolio within some defined feasible set; that is to say, an optimisation is performed. The key issues with such approaches are the following:

- Are the relationships (e.g., correlations or other measures of interdependence) used in the optimisation relevant for identifying worst-case scenarios?
- Is the algorithm capable of identifying the globally worst-case outcome in the feasible set?
- Does the feasible set include implausible outcomes, and are those outcomes represented disproportionately in the optimal results?

Two approaches are discussed below, namely, factor-push and maximum loss.

III.A.4.8.1 Factor-Push Stress Tests

This type of stress test is so named because it involves ‘pushing’ each individual market risk factor in the direction that results in a loss for the portfolio. Construction is straightforward.

1. A push magnitude, m , is selected; it can be stated as a number of standard deviations. For example, each market risk factor, r , may be pushed four standard deviations. The magnitude chosen is, ultimately, subjective. It may be chosen with reference to (an average of) observed movements in market prices during some significant historical event. Or it may be chosen to correspond to some quantile of returns based on an assumed distribution (or perhaps an empirical distribution quantile). If you are setting the push magnitude using standard deviations or quantiles, the period over which these are estimated must be chosen as well. If you are using unconditional estimates, then one year of daily return history is good, if available. When using a methodology for estimating volatility conditionally, such as GARCH, then it is good to use as much return history as there is reliable data.
2. The portfolio, P , is revalued twice by applying shocks, s , to a single market risk factor: once applying a shock $s^+ = (+1 \times m)$, and once applying a shock $s^- = (-1 \times m)$.
3. The two portfolio revaluations are compared and the shock resulting in the lower portfolio value is adopted for the stress test.
4. Steps 2 and 3 are repeated for each of the N market risk factors affecting the portfolio.
5. The portfolio is revalued once more, this time simultaneously applying the shocks selected in the prior steps for each risk factor.

Consider the following example. A portfolio consists of a long position of 1000 shares in IBM and a short position of 1700 shares in GM. On 23 February 2004 the closing prices of the two equities were \$95.96 and \$47.51, respectively. Their respective daily standard deviations of return (based on one year of closing prices) were 0.005955 and 0.006972, respectively. Set the push magnitude to 6 (more on this below). The shock for IBM will be +1, because the position is long

and the return is linear in the price change. The shock for GE will be -1 , by analogous reasoning. The factor-push portfolio stress loss is then \$6807.09.

The selection of the push magnitude is subjective and ignores correlations. Note that the largest one-day move downward in IBM in the five years ending 23 February 2004 is -9.31% , or about 15.63 times the estimated standard deviation of IBM. The third largest move downward is still 7.78 times the standard deviation. The largest one-day move downward in GE in the same period is -6.31% , or 9.06 times the estimated standard deviation of GE. The third largest move down in GE is 5.35 times the standard deviation.

If we assume, as in a historical simulation, that any of the one-day IBM–GE return pairs observed over that five year period were equally possible when looking forward one day, then the potential worst-case portfolio loss would be \$22,175.00, much bigger than the factor-push loss. Of course, the factor-push loss can be made arbitrarily large by increasing the push magnitude. However, the goal is not to create an arbitrarily large loss, but rather to estimate a plausible, if unlikely, loss.

Small perturbations of the push magnitudes, if applied to individual positions, can result in greater estimated losses. For example, if the magnitude applied to IBM is reduced to 5.99 and the magnitude applied to GE is increased to 6.01, the factor-push loss will increase (slightly) because GE has the greater estimated standard deviation. This fact highlights the implicit assumption here that the marginal return distributions have the same shape and thus the marginal probabilities of the moves are equal. It may be better to select push-magnitudes position by position, based on the volatility of each factor.

A simple thought experiment also illustrates that the factor-push loss need not be greater than the losses from all unambiguously smaller moves, if some of the portfolio positions have returns that are nonlinear functions of the market risk factors. A long option straddle has its least payout associated with small moves in the underlying market factor. Since the factor-push method only ‘searches’ among large moves for the worst outcomes, the factor-push method will be less useful in such cases.

This type of stress test has a variety of other drawbacks as well. Most importantly, these stress tests ignore correlations among risk factors. As a result, the specification of the factor changes employed in the stress test may imply highly implausible market dynamics. For example, neighbouring tenor points of a given yield curve are generally very highly positively correlated. Consider a situation where the push factor is 4 standard deviations, and the stress shock at the

three-month point on a given yield curve is -1 , while at the adjacent six-month point the shock is $+1$. Such a scenario fails the ‘laugh test’. Also, cross effects in instruments whose values are affected by multiple risk factors are ignored. For example, for an option on an equity, the option delta is itself related to the volatility of the equity. Thus the correct choice of shock, $+1$ or -1 , for the equity price is dependent on the choice of shock for the volatility. Choosing the push factor based on the standard deviation, or even the VaR, may defeat the purpose of the test. If it is the case that extreme events are, in effect, observations from a distribution that is distinct from that which is experienced most of the time, there may be no obvious choice of push factor for the portfolio that ‘maps’ from the estimated standard deviations to a stress environment. The particular set of shocks that are chosen can change with every run of the stress test. This makes it more difficult to communicate intuition about the nature of the shocks that generated the observed stress-test result.

III.A.4.8.2 Maximum Loss

A maximum loss scenario is defined by the set of changes in market risk factors that results in the greatest portfolio loss, subject to some feasibility constraint on the allowable changes in market risk factors (see Studer, 1995, 1997). The constraint is necessary because potential portfolio losses need not be bounded, and thus, absent a constraint, the result may lack plausibility. Note that the maximum loss scenario is dependent on the structure of the portfolio, and not related to any particular economic environment, except as an environment is reflected in the statistical description of market factor returns employed in the analysis. Thus it is well adapted to answer a question such as just how bad can things plausibly get.¹¹

One natural method of constraining the set of feasible risk factors is to consider only those scenarios that have a likelihood in excess of some small probability. Being able to associate a probability with a stress-test loss is potentially very useful. To proceed down this path it is necessary to be able to state joint probabilities of specific sets of risk factor changes, which in turn requires that some statements be made about the co-movements of risk factors. Those statements will be based on either historical or simulated data and can be conditional (e.g., stressed) or unconditional (‘normal’).

Note that, in general, to apply this approach, it is necessary to search through the entire space of risk factor changes with joint probability less than or equal to the plausibility constraint. With

¹¹ Boudoukh *et al.* (1995) propose a risk measure they call the worst-case scenario (WCS). More akin to conditional VaR than stress testing, WCS is the expected value of the distribution of maximum loss over some time horizon, for example, the expected worst-day portfolio P&L over a month. It is also related to the extreme-value theory method discussed below.

positions whose values are nonlinear functions of risk factors in the portfolio, the maximum loss need not occur at the limits of the allowable risk factor changes. Breuer and Krenn (2000) suggest Monte Carlo simulation (or some quasi-random method) as a natural method to use in this regard. A naïve approach would be to run the Monte Carlo for a predetermined sample size, throw out any draw of (an n -tuple of) factor returns with a corresponding theoretical probability less than the plausibility constraint, and pick from the remaining sampled (n -tuples of) factor returns the one with the largest associated portfolio loss.

Practically, however, too many simulated risk factor vectors would be required to attain a high level of confidence that the maximum loss had been identified using this approach. Quasi-random methods may ensure a more efficient sampling of the feasible return space. However, since a complex portfolio may have many local return minima and possibly many discontinuities in the portfolio return function, even quasi-random methods may not make the search for that maximum loss tractable. To employ this approach in a practical manner, it may be necessary to develop a ‘smart’ search algorithm that, based on the qualities of the positions in the portfolio, makes guesses as to which parts of the return space can be ignored.¹²

III.A.4.9 Extreme-Value Theory as a Stress-Testing Method

Extreme-value theory (EVT) is based on limit laws which apply to the extreme observations in a sample. These laws allow parametric estimation of high quantiles of loss (negative return) distributions without making any (substantial) assumptions about the shape of the return distribution as a whole. The application to stress testing is immediate (sort of). Firstly, it must be assumed that the historical data employed in the estimation of the tail are representative (i.e., are drawn from the distribution relevant for examining the stress event). Second, the size of the historical data sample must be large enough to yield enough ‘tail’ observations to get good estimates of the tail distribution’s parameters.

Two flavours of EVT have been employed in looking at risk measurement in finance. The first is called the ‘block maxima’ approach. An example is the set of yearly maxima of negative daily returns on the S&P500 over some period. The second method is the ‘peak-over-threshold’ approach, illustrated by the greatest z (i.e., some natural number of) negative daily returns on the S&P500 over the same period. It is apparent that the two methods have somewhat different definitions of extreme events. The block maxima approach may be better suited to estimation of stress losses and the peak-over-threshold approach may be better suited to VaR estimation. If we take the block to be a year, then we can use this approach to find the loss that is expected to be

¹² Gonzalez-Rivera (2003) develops an interesting related approach.

exceeded once in every k years. The details of this approach are beyond the scope of this chapter. For an application of the approach to stress testing, see Cotter (2000).

III.A.4.10 Summary and Conclusions

The main points of this chapter are as follows:

- ❖ Stress testing is perceived as being a useful supplement to value-at-risk because value-at-risk does not convey complete information about the risk in a portfolio.
- ❖ Stress tests should be part of a rational economic approach to decision making. However, in practice, most uses of stress results are somewhat *ad hoc*.
- ❖ Useful stress tests must represent plausible, if unlikely, factor changes. Creating useful stress tests requires detailed consideration of the potential behaviour of the market risk factors, individually and jointly, and an understanding of structural relationships among risk factors (e.g., no arbitrage requirements and spread relationships).
- ❖ Stress-test methods generally fall into three categories, historical scenarios, hypothetical scenarios, and algorithms.
 - Historical scenarios seek to re-create a particular economic environment from the past,
 - Hypothetical scenarios can represent a complete, but not yet experienced, economic story or simply a set of *ad hoc* factor movements.
 - Algorithms attempt to systematically identify the set of factor changes (within some bounds) that give the worst-case portfolio loss. In the case of the factor-push method, the result may not represent a plausible economic story.
- ❖ There is no best or right type of stress test. The context in which the results will be used should determine the approach to be taken in stress testing.

Further Reading

Most of the papers listed here and in the References following may be found on www.GloriaMundi.org

Basel Committee on Banking Supervision (1999) Recommendations for public disclosure of trading and derivatives activities of banks and securities firms. Mimeo (October).

Bouyé, E (2002) Multivariate extremes at work for portfolio risk management. Working Paper (January).

Bouyé, E, Durrleman, V, Nikeghbali, A, Riboulet, G, and Roncalli, T (2000) Copulas for finance: A reading guide and some applications. Working Paper (July).

Bouyé, E, Durrleman, V, Nikeghbali, A, Riboulet, G, and Roncalli, T (2001) Copulas: An open field for risk management. Working Paper (March).

- Čihák, Martin (2004a) Designing stress tests for the Czech banking system. Czech National Bank, mimeo (March).
- Čihák, Martin (2004b) Stress testing: A review of key concepts. Czech National Bank, mimeo (February).
- Costinot, A, Riboulet, G, and Roncalli, T (2000) Stress testing et théorie des valeurs extrêmes: Une vision quantifiée du risque extrême. Working paper, Credit Lyonnais (September).
- Hong Kong Monetary Authority (2003) Stress testing. Supervisory Policy Manual IC-5 (February).
- International Association of Insurance Supervisors (2003) Stress testing by insurers. Guidance paper (October).
- Jouanin, J-F, Riboulet, G, and Roncalli, T (2004) Financial applications of copula functions. In G Szego (ed.), *Risk Measures for the 21st Century*. New York: Wiley.
- Oesterreichische Nationalbank (1999) Guidelines on market risk, Volume 5: Stress testing. Mimeo.
- Wee, L-S, and Lee, J (1999) Integrating stress testing with risk management. *Bank Accounting & Finance* (Spring), pp. 7–19.

References

- Basel Committee on Banking Supervision (1994) Risk management guidelines for derivatives. Mimeo (July).
- Basel Committee on Banking Supervision (1995) An internal model-based approach to market risk capital requirements. Mimeo (April).
- Basel Committee on Banking Supervision (1996) Amendment to the capital accord to incorporate market risks. Mimeo (January).
- Basel Committee on Banking Supervision (2002) Basel Committee reaches agreement on new capital accord issues. Press release (10 July).
- Berkowitz, J (1999) A coherent framework for stress testing. *Journal of Risk*, 2 (Winter), pp. 1–11.
- Blaschke, W, Jones, M T, Majnoni, G, and Martinez Peria, S (2001) Stress testing of financial systems: An overview of issues, methodologies and FSAP experiences. Mimeo (June).
- Borio, C (2000) Market liquidity and stress: Selected issues and policy implications. *BIS Quarterly Review* (November), pp. 38–51.
- Boudoukh, J, Richardson, M, and Whitelaw, R (1995) Expect the worst. *Risk*, 8(9), pp. 100–101.
- Boyer, B, Gibson, M, and Loretan, M (1999) Pitfalls in tests for changes in correlations. Working paper, Federal Reserve Board.
- Breuer, T, and Krenn, G (2000) Identifying stress test scenarios. Working paper.

- Cherubini, U, and Della Longa, G (1999) Stress testing techniques and value-at-risk measures: A unified approach. *Rivista di Matematica per le Scienze Economiche e Sociali*, 22(1/2), pp. 77–99.
- Committee on the Global Financial System (2001) A survey of stress tests and current practice at major financial institutions. Mimeo (April).
- Comptroller of the Currency (1993) Banking Circular 277: Risk management of financial derivatives. Mimeo (October).
- Corsetti, G, Pericoli, M, and Sbracia, M (2002) Some contagion, some interdependence. Working paper, Yale University.
- Cotter, J (2000) Crash and boom statistics for global equity markets. Working paper, University College Dublin.
- Culp, C and Miller, M (eds) (1999) *Corporate Hedging in Theory and Practice: Lessons from Metallgesellschaft*. London: Risk Publications.
- Davis, E P (2003) Towards a typology for systematic financial instability. Working paper, Brunel University (November).
- Derivatives Policy Group (1995) Framework for voluntary oversight. Mimeo (March).
- Dungey, M, and Zhumabekova, D (2001) Testing for contagion using correlations. Working paper, Australian National University.
- Fender, I, and Gibson, M (2001a) The BIS census on stress tests. *Risk*, 14(5), pp. 50–52.
- Fender, I, and Gibson, M (2001b) Stress testing in practice: A survey of 43 major financial institutions. *BIS Quarterly Review* (June), pp. 58–62.
- Fender, I, Gibson, M, and Mosser, P (2001) An international survey of stress tests. *Current Issues in Economics and Finance*, 7(10), pp. 1–6.
- Finger, C (1997) A methodology to stress correlations. *RiskMetrics Monitor*, 3-12.
- Forbes, K, and Rigobon, R (2002) No contagion, only interdependence. *Journal of Finance*, 43.
- Frye, J (1997) Principals of risk: Finding value-at-risk through factor-based interest rate scenarios. In S Grayling (ed.), *Understanding and Applying Value at Risk*. London: Risk Books.
- Gay, G, Kim, J, and Nam, J (1999) The case of the SK Securities and J.P. Morgan swap: Lessons in VaR frailty. *Derivatives Quarterly*, 5(Spring), 13–26.
- Gilbert, C L (1996) Manipulation of metals futures: Lessons from Sumitomo. Working paper, University of London (November).
- Global Derivatives Study Group (1993) *Derivatives: Practices and Principles*. Washington, DC: Group of Thirty
- Gonzalez-Rivera, G (2003) Value in stress: A coherent approach to stress testing. *Journal of Fixed Income* (September), pp. 7–18.
- Greenspan, A (2003) Corporate governance. Speech at the Conference on Bank Structure and Competition (May).

- Hickman, A, and Jameson, R (2001) Benchmarking the US attack crisis. ERisk.com (September).
- Higham, N J (2002) Computing the nearest correlation matrix – a problem from finance. *IMA Journal of Numerical Analysis*, 22, pp. 329–343.
- Hoggarth, G, and Whitley, J (2003) Assessing the strength of UK banks through macroeconomic stress tests. *Financial Stability Review*, 3(6), pp. 91–103.
- International Monetary Fund (2003) Analytical tools of the FSAP. Mimeo (February).
- Jeffery, C (2003) The ultimate stress test: Modeling the next liquidity crisis. *Risk* (November).
- Jorion, P (1995) *Big Bets Gone Bad*. Amsterdam: Elsevier Science.
- Jorion, P (2002) Risk management in the aftermath of September 11. Working paper, University of California-Irvine (April).
- Kim, J, and Finger, C (2000) A stress test to incorporate correlation breakdown. *Journal of Risk*, 2(1).
- Kupiec, Paul (1998) Stress testing in a value at risk framework. *Journal of Derivatives*, 6, pp. 7–24.
- Loretan, M, and English, W (2001) Evaluating correlation breakdown during periods of market volatility. Working paper.
- Malz, A, and Mina, J (2001) Risk management in the aftermath of the terrorist attack. Working paper, RiskMetrics Group (September).
- MacKenzie, D (2003) An equation and its words: Bricolage, exemplars, disunity and performativity in financial economics. Working paper.
- Overdahl, J and Schachter, B (1995) Derivatives regulation and financial management: Lessons from Gibsons Greetings. *Financial Management*, 24, pp. 68–78.
- President's Working Group on Financial Markets (1999) Hedge funds, leverage, and the lessons of Long Term Capital Management. Mimeo.
- Rebonato, R, and Jäckel, P (1999) The most general methodology to create a valid correlation matrix for risk management and option pricing purposes. *Journal of Risk*, 2 (2).
- Schachter, B (1998a) The value of stress testing in market risk management. *Derivatives Risk Management* (March).
- Schachter, B (1998b) Move over VaR. *The Financial Survey* (May), 12–14.
- Schachter, B (2000a) How well can stress testing complement VaR? In, T. Haight, *Derivatives Risk Management*. (Arlington, Virginia) A. S. Pratt & Sons.
- Schachter, B (2000b) Stringent stress tests. *Risk*, 13, (December), pp. S22–S24.
- Securities and Futures Authority Limited (1995) Value-at-risk models. Board Notice 254, 26 May.
- Shaw, J (1997) Beyond VaR and stress testing. In *VaR: Understanding and Applying Value-at-Risk*. New York: Risk Publications, pp. 211–224.

Studer, G (1995) Value at risk and maximum loss optimization. Working paper, ETHZ (December).

Studer, G (1997) Maximum loss for measurement of market risk. Doctoral thesis, Swiss Federal Institute of Technology.

Turkay, S, Epperlein, E, and Nicos, C (2003) Correlation stress testing for value-at-risk. *Journal of Risk*, 5(4), pp. 75–89.

Zangari, P (1997a) Exploratory stress-scenario analysis with applications to EMU. *RiskMetrics Monitor*, Special Edition, pp. 31–54.

Zangari, P (1997b) Catering for an event. *Risk* (July), pp. 34–36.

III.B.2 Foundations of Credit Risk Modelling

Philipp Schönbucher¹

III.B.2.1 Introduction

This chapter introduces the three basic components of a credit loss: the exposure, the default probability and the recovery rate. We define each of these as *random* processes, that is to say, their future values are not known. Instead, their values are represented as a probability distribution of a random variable. The credit loss distribution is then defined as a product of these three distributions. The credit loss distribution for a large portfolio of credits can become quite complex and we shall see in Chapter III.B.5 that many advanced techniques are available for portfolio modelling. In this chapter we merely aim to provide an introduction to this distribution.

Section III.B.2.2 makes the definition of default risk precise, and Sections III.B.2.3 and III.B.2.4 introduce the three basic processes and the credit loss distribution. Section III.B.2.5 makes a careful distinction between expected and unexpected loss, and Section III.B.2.6 gives a detailed discussion of recovery rates for a portfolio of credits. Section III.B.2.7 summarizes and concludes.

III.B.2.2 What is Default Risk?

Default risk is the risk that a counterparty does not honour his obligations. Such an obligation may be a payment obligation, but it is also a default if a supplier does not deliver the parts he promised to deliver, or if a contractor does not render the services that he promised. In this wide context, default risk is everywhere, and there is no transaction that does not involve the risk that one of the two parties involved does not deliver.

For a financial institution the largest and most important component of default risk refers to payment obligations such as loans, bonds, and payments arising from over-the-counter derivatives transactions. This risk of a payment default, in particular when it refers to loans and bonds, is called *credit risk*. We distinguish:

- *Default*: An obligation is not honoured.
- *Payment default*: An obligor does not make a payment when it is due. This can be:
 - *Repudiation*: Refusal to accept a claim as valid.

¹ D-MATH, ETH Zürich, Rämistrasse 101, 8092 Zürich

- *Moratorium*: Declaration to stop all payments for some period of time. Usually only sovereigns can afford to do this.
- *Credit default*: Payment default on borrowed money (loans and bonds).
- *Insolvency*: Inability to pay (even if only temporary).
- *Bankruptcy*: The start of a formal legal procedure to ensure fair treatment of all creditors of a defaulted obligor.

For instance, an obligor may be:

- in default but not in payment default (e.g. if he defaults on a non-financial obligation),
- in payment default but not insolvent (e.g. if he could pay, but chooses not to), and
- in default but not in bankruptcy (e.g. if the bankruptcy procedures have not been started (yet) or if there are no bankruptcy procedures, e.g. if the obligor is a sovereign).

Strictly speaking, default risk is tied to the obligation it refers to, and *a priori* there is no reason why an obligor should not default on one obligation and honour another. Fortunately, in most cases we can rely on the existence of a functioning legal system which ensures that such selective defaults are not possible. The creditor of the defaulted obligation has the right to go to a court which (eventually) will force the obligor to honour his obligation (if this is possible) or – if the obligor is generally unable to do so – instigate a formal bankruptcy procedure against him. The aim of this procedure is an orderly and fair settlement of the creditors' claims and possibly also other social priorities such as the preservation of jobs. The details of the bankruptcy procedure depend on the applicable local bankruptcy law and will vary across countries. Thus, the existence of a bankruptcy code allows us to speak of the credit risk of an obligor and not just of an individual obligation, which we will do in the following.

III.B.2.3 Exposure, Default and Recovery Processes

To analyse the components of default risk in more detail we now need to introduce some notation. We consider a set of I obligors indexed with $i = 1, \dots, I$, and we call τ_i the time of default of the i^{th} obligor. The following three processes describe the credit risk of obligor i :

- $N_i(t)$: *The default indicator process*. At time t , $N_i(t)$ takes the value one if the default of obligor i has occurred by that time, and zero if the obligor is still alive at time t . Obviously knowledge of the full path of the default indicator process is equivalent to knowledge of the exact time of default of the obligor, but frequently we only have partial knowledge (i.e. the obligor has not defaulted so far), or we are only interested in a partial event (i.e. default before the maturity of a loan).

- $E_i(t)$: *The exposure process*. The exposure at default (EAD) to obligor i at time t is the total amount of the payment obligations of obligor i at time t which would enter the bankruptcy proceedings if a default occurred at time t . The exposure process will be covered in detail in Chapter III.B.3.
- $L_i(t)$: *The loss given default (LGD)* of obligor i , given default at time t . The LGD usually takes values between zero and one, and $R_i(t) = 1 - L_i(t)$ is known as the *recovery rate* of the obligor. The LGD is often less than one, reflecting the fact that some proportion of the exposure at default may be recovered in bankruptcy proceedings. We will discuss recovery rates in more detail in Section III.B.2.6 .

Let $p_i(T)$ be the individual *probability of default* (PD) of obligor i until some time horizon T . For instance, if the one-year probability of default is $p_i(1) = 0.01\%$, this means that there is a 1/10,000 chance that the obligor will default at some point in time over the next year.

Using these processes, the losses due to defaults of obligor i before time T can be represented as follows:

$$\text{Default loss} = \text{Default arrival} \times \text{EAD} \times \text{LGD}$$

or, in our mathematical notation where τ denotes the time of default of the obligor,

$$D_i(T) = N_i(T) \times E_i(\tau) \times L_i(\tau). \quad (\text{III.B.2.1})$$

For a fixed time-horizon T , the default indicator $N_i(T)$ is a binary (0/1) variable which allows one to capture the default arrival risk, that is, the risk whether a default occurs at all. A fixed time-horizon (typically one year) is a common point of view in credit risk management, but in many cases this is not sufficient and the timing risk of defaults must also be considered.

III.B.2.4 The Credit Loss Distribution

For a bank, individual credit defaults of obligors are not unusual events and – although painful and inconvenient – these events are part of the normal course of business. If the exposure is not too large it can be buffered using normal operating cash flows. But when multiple defaults occur simultaneously (or within a short time span) this can threaten the existence of a financial institution. Thus, a major task of the credit risk manager is to measure and control the risk of losses from a whole portfolio of credits. For instance, suppose the credit losses are correlated, so that a default from one obligor makes default of other obligors more likely. Then there is a *concentration risk* in the portfolio that needs to be managed.

Using the definition of individual default losses (III.B.2.1), the portfolio loss $D(T)$ at time T is defined as the sum of the individual credit losses $D_i(T)$, summed over all obligors $i = 1, \dots, I$:

$$D(T) = \sum_{i=1}^I D_i(T) = \sum_{i=1}^I N_i(T) \cdot E_i(\tau) \cdot L_i(\tau). \quad (\text{III.B.2.2})$$

The individual default losses $D_i(T)$ are unknown in advance. Thus the full portfolio loss $D(T)$ is a random variable. The portfolio's *credit loss distribution* is the probability distribution of this random variable (see Chapter II.E).

$$F(x) := P[D(T) \leq x]. \quad (\text{III.B.2.3})$$

An important ingredient of the distribution of $D(T)$ is the dependency between the individual losses. In Chapter III.B.4 some popular models are presented which show how default correlation can be modelled. Here we only mention two important points:

- (a) Assuming independence between individual default losses almost always leads to a gross underestimation of the portfolio's credit risk.
- (b) The default correlation parameters typically have a strong influence on the tail of the loss distribution – and thus on the value at risk (VaR).

Figure III.B.2.1: Density function of the portfolio loss of a typical loan portfolio. Mean loss (expected loss), 99% VaR, and 99.9% VaR are shown as vertical lines

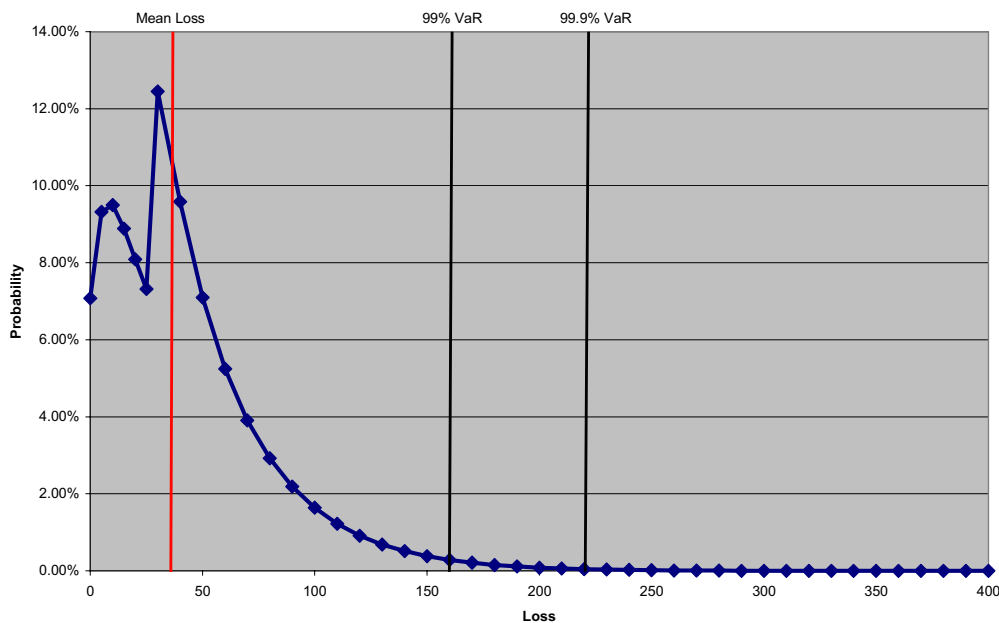


Figure III.B.2.1 shows the density function of the credit loss distribution of a typical credit portfolio using a hypothetical portfolio of 100 obligors with 10m exposure each (i.e. 1bn total portfolio volume). We have used a CreditMetrics type of model (see Section III.B.5.3) with individually varying unconditional default probabilities, 50% assumed recovery rate and a constant asset correlation of 20%.

Credit portfolio loss distributions have several features that distinguish them from the ‘profit and loss’ distributions (or returns distributions) from market variables such as equities, interest rates or FX markets:

- The distribution is *not symmetrical*. The ‘upside’ is limited (the best possible case is a credit loss of zero), while the downside can become extremely large.
- The distribution is *highly skewed*. Most of the probability mass is concentrated around the low loss events (e.g. losses between 0 and 50 in Figure III.B.2.1). These are also the events that we are most likely to observe in historical data sets. In the example of Figure III.B.2.1, losses will be less than 60m in 80% of the cases.
- The distribution has *heavy tails*. This means that the probability of large losses decreases very slowly, and VaR quantiles are quite far out in the tail of the distribution. This can also be seen from Figure III.B.2.1.

III.B.2.5 Expected and Unexpected Loss

The ‘standard scenario’ which people intuitively expect to happen when they consider the default risk of an obligor is ‘no default’, that is, no loss. This is indeed the most likely scenario if we consider each obligor individually (unless we consider an obligor of extremely low credit quality). This scenario is also still used quite frequently for accounting purposes: a loan or bond is booked at its notional value (essentially assuming zero loss), and only if it gets into distress is it depreciated. In some institutions return on capital is still (incorrectly) calculated this way.

Unfortunately, this is one of the cases where naive intuition can lead us astray: the ‘standard scenario’ is not the mathematical expectation of the loss on the individual obligor. If we assume that exposure E_i and loss given default L_i are known and constant, the *expected loss* is

$$E[D_i(T)] = p_i(T) \cdot E_i \cdot L_i \neq 0, \quad (\text{III.B.2.4})$$

where $p_i(T)$ is the default probability of the obligor.

At the level of individual obligors in isolation, the concept of expected loss may be counter-intuitive at first because we will never observe a realisation of the expected loss: either the obligor survives (then the realised loss will be zero) or the obligor defaults (then we will have a realised loss which is much larger than the expected loss).

There is a related trick question. Next time you go out, offer to buy your friend a drink if the next person entering the bar does not have an above average number of legs. Of course your friend will have to buy you a drink if the person does indeed have an above average number of legs. You will win the bet if the next person has two legs: the average number of legs per person in the population must be slightly less than two because there are some unfortunate people who have lost one or both legs (but there are no people with more than two legs).

The same idea applies to credit obligors: most obligors will perform better than expected (they will not default), but there are some who perform significantly worse than expected. But nobody will perform exactly as expected.

Typically, the expected loss is small (because p_i will be small) but it is positive. These small errors will accumulate when we consider a portfolio of many obligors. In a portfolio of 1000 obligors we may no longer assume that none of these obligors will default. Even if each of the obligors has a default probability of only 1%, we will have to expect 10 defaults. In Figure III.B.2.1 the level of the expected loss of the portfolio is shown by the first (leftmost) vertical line; it is at a level of about 35m, that is, 3.5% of the portfolio's notional amount.

Expected loss is an important concept when it comes to performance measurement, in particular in connection with risk-adjusted return on capital (RAROC) calculations (see Chapter III.0). When a loan's expected gain (in terms of excess earnings over funding and administration costs) is not sufficient to cover the expected loss on this loan, then the transaction should not be undertaken. Suffering the expected loss (in particular, on a portfolio) is not bad luck: it is what you should *expect* to happen. Consequently, the expected loss should be covered from the portfolio's earnings. It should *not* require capital reserves or the intervention of risk management.

The expected loss on a portfolio is the sum of the expected losses of the individual obligors:²

$$E[D(T)] = \sum_{i=1}^I E[D_i(T)]. \quad (\text{III.B.2.5})$$

² This follows from the property of the expectation operator: $E(X + Y) = E(X) + E(Y)$ – see Chapter II.E.

However, this simple summation property will *not* hold for the *unexpected loss*! *Unexpected loss* is usually defined with respect to a VaR quantile and the probability distribution of the portfolio's loss. Let us assume that $D_{99\%}$ is the portfolio's 99% VaR quantile, that is,

$$P[D(T) \leq D_{99\%}] = 99\%.$$

Then the unexpected loss of the portfolio at a VaR quantile of 99% is defined as the difference between the 99% quantile level and the expected loss of the portfolio:

$$UEL = D_{99\%} - E[D(T)] \tag{III.B.2.6}$$

If another risk measure such as conditional VaR is used in place of VaR, then (III.B.2.6) is easily extended. We define unexpected loss in these situations by replacing $D_{99\%}$ with the general risk measure. More details on some alternative risk measures are given in Section III.A.3.5.2.

The term 'unexpected loss' may be confusing at first, because it does not concern losses that were unexpected, but only something like a worst-case scenario. Intuitively, one might define unexpected loss as the amount by which the portfolio's credit loss turns out, in the end, to exceed the originally expected loss:

$$\max\{D(T) - E[D(T)], 0\}. \tag{III.B.2.7}$$

Here, we will use unexpected loss in the sense of equation (III.B.2.6), and not in the sense of (III.B.2.7).

In Figure III.B.2.1, the 99% and 99.9% VaR loss quantiles were shown with two vertical lines intersecting the tail of the loss distribution. The 99% VaR level is at approximately 160m, which yields an unexpected loss of $160\text{m} - 35\text{m} = 125\text{m}$. The 99.9% VaR level is at approximately 220m, with an unexpected loss of 185m. It is no coincidence that, even at a very high VaR quantile of 99.9%, the unexpected loss is still much less than the maximum possible loss of 1bn that is suffered when the total portfolio defaults with zero recovery. This effect stems from the partial diversification, which is still present in the portfolio despite an asset correlation of 20%.

As opposed to expected loss, the unexpected loss is not additive in the exposures. If, for example, we assume that with zero recovery and unit exposure each obligor defaults with a probability of 3%, then each obligor's individual 99% VaR will be 1, its total exposure. But the 99% VaR of a large portfolio of such obligors will not be the total exposure of the portfolio (unless we have the extreme case of perfect dependency between all obligors).

The unexpected loss is frequently used to determine the capital reserves that have to be held against the credit risk of the portfolio. It is not economically viable to hold full reserves against total loss of the portfolio, but reserving against unexpected loss at a sufficiently high quantile is viable and effective if it is done centrally under exploitation of all diversification effects. Thus, coverage of the (linear) expected loss is in the domain of responsibility of the business lines, but the management of the (highly non-linear) unexpected loss is usually a task for a centralised risk management department. The risk management department then makes appropriate risk charges to the business lines.

A stylised *capital allocation* procedure is as follows:

1. Fix a VaR quantile for credit losses (usually 99% or 99.9%). This quantile should reflect the institution's desired survival probability due to credit losses (this probability can be derived from its targeted credit rating). This is a management decision that has to be made at the top level.
2. Determine the portfolio's expected loss.
3. Determine the unexpected loss of the portfolio according to (III.B.2.6).
4. Allocate risk capital to the portfolio to the amount of the unexpected loss.
5. Split up the portfolio's risk capital over the individual components of the portfolio according to their risk capital contributions.

Losses in the portfolio up to the amount of the expected loss will have to be borne by the individual business lines (because these losses are economic losses), but any losses that exceed the expected loss will hit the risk capital reserves. Should these reserves not suffice to cover all losses, the bank itself will have to default. But by setting the original VaR level, the probability of this event can be controlled.

III.B.2.6 Recovery Rates

As we can see from equations (III.B.2.1) and (III.B.2.4), the recovery rate (or the LGD) of an obligor is as important in determining default losses or expected default losses as the default probability. Nevertheless, research into recovery rates has been neglected for a long time, and most focus has been put on default events and default probabilities. This is partly due to the fact that data on recovery rates are much more fragmented and unreliable than data on default events.

While we may expect that obligors will do their best to avoid bankruptcy in almost any legal environment (which will make default arrivals largely independent of the legal framework), recovery rates are closely tied to the legal bankruptcy procedure which is entered upon the default of the obligor.

In a typical bankruptcy procedure all creditors register their legal claim amounts with the bankruptcy court. There is a well-defined procedure to determine these legal claim amounts which does not necessarily reflect the market value of the claim: for example, for loans or bonds only the notional amount (and any interest payments which are currently due) is considered, but not the actual market value of the bond or the value of the future coupon payments. These values may be significant if interest rates have moved since the issuance of the bond. According to the ISDA standard definitions, the legal claim amount for over-the-counter derivatives is usually the current replacement value of the contract, where it is assumed that the counterparty has the same rating as the defaulted counterparty's pre-default rating.

Claims are then grouped by priority class (collateralised, senior, junior, etc.). At this point, the bankruptcy procedures start to diverge significantly: some procedures aim to find a way to restructure the bankrupt obligor and to enable him to become profitable again (e.g. Chapter 11 in the USA), while others aim is to liquidate the obligor's business and use the proceeds to pay off the debts (e.g. Chapter 7 bankruptcy in the USA). The final outcome for the creditors usually depends first on the choice of bankruptcy procedure, and then on complicated negotiations between many parties. We will have to ignore these effects here, and just warn the reader that, because of the strong dependency on procedural details, recovery rates are not easily comparable across countries.

Generally, the outcome of any bankruptcy procedure will be a settlement in which creditors of the same legal claim amount and the same priority class will be treated identically, with secured creditors having the first claim on their collateral and the rest of the firm's assets, unsecured creditors come next, and then stockholders have the last claim and may not receive anything.

If the creditor's claims are settled in cash, the definition of the recovery rate is relatively straightforward: if each dollar of legal claim amount receives a 40 cent cash settlement, then the recovery rate is 40% and the LGD is $1 - 40\% = 60\%$. The only problem is to decide whether the payment should be discounted back to the actual date of default (in particular, for large obligors, bankruptcy procedures can easily take several years, and the usual answer is 'yes'), and whether legal costs should be subtracted from it (again 'yes').

Unfortunately, cash settlements tend to be rare. Much more frequently (in particular, if the obligor is restructured and not liquidated) the settlement is partly cash, and partly some other type of security such as equity, preferred equity or restructured debt of the reorganised obligor. In this case, determining the value of the settlement is often close to impossible, in particular if

the obligor was de-listed from the stock exchange in the course of the bankruptcy. For these reasons, recovery rates are generally defined without reference to the final settlement.

Definition 1 (Market Value Recovery). *The recovery rate is the market value per unit of legal claim amount of defaulted debt, some short time (e.g. 1 or 3 months) after default.*

This definition also coincides with the way recovery rates are determined in credit default swaps with cash settlement (see Chapter I.B.6). It generally applies to larger obligors (e.g. obligors rated by public rating agencies).

When smaller obligors are concerned (e.g. retail obligors and/or small and medium-sized enterprises) there will be no market price for distressed debt and we will have to attach a direct valuation to the final default settlement, and define the recovery rate as: follows

Definition 2 (Settlement Value Recovery). *The recovery rate is the value of the default settlement per unit of legal claim, discounted back to the date of default and after subtracting legal and administrative costs.*

According to the discussion above, we expect the following factors to directly influence the recovery rates of defaulted debt:

- collateral;
- the legal priority class of the claim;
- the legislature in which the bankruptcy takes place (the UK tends to be a rather creditor-friendly legislature, while France and the USA are more obligor-friendly and thus have lower recovery rates).

Some other, less easily observed factors turned out to be significant in empirical investigations (see, for example, Altman *et al.*, 2001; Gupton *et al.*, 2001; Van de Castle *et al.*, 2001; Renault and Scaillet, 2004):

- *The industry group of the obligor.* Financial institutions tend to have significantly different (higher) recovery rates than industrial obligors. The less capital intensive the obligor's business, the less substance there is to liquidate in the event of a bankruptcy. For instance, 'dotcom' companies tended to have recoveries close to zero.
- *The obligor's rating prior to default.* An obligor who has spent much time close to default usually has fewer assets to liquidate to pay off the creditors than an obligor who defaulted quite suddenly from a high rating class.

- *The average rating of the other obligors in the industry group, and the business cycle.* This affects the liquidation value of the obligor’s business and/or the value and viability of a restructured firm. Recoveries tend to be lower in recessions and in industry groups that are in cyclical downswings or which have large overcapacities.

Despite these empirical findings, it turns out that it is virtually impossible to predict a recovery rate of an obligor with much certainty. The margins of error are very large indeed. Table III.B.2.1 shows estimated recovery rates and their standard errors for US corporate debt of different seniority classes. It can be seen that the average standard error is often of the same order of magnitude as the mean of the recovery rate.

Table III.B.2.1: Recovery rates by seniority of claim

Seniority	Observations	Mean (%)	Standard deviation (%)
Senior secured	82	56.31	23.61
Senior unsecured	225	46.74	25.57
Subordinated	174	35.35	24.64
Junior subordinated	142	35.03	22.09
Total	623	42.15	25.42

Source: Renault and Scaillet (2004) Data set: S&P, 1981–1999, US Corporates

From a risk management point of view the large prediction errors in the recovery rates would not be too serious if we could at least hope that our estimation errors will cancel out on average over several defaults. Unfortunately, the systematic dependence of recoveries on the business cycle destroys this hope. Recoveries depend on a common factor and thus they will not diversify away. In particular, we will get hit twice in a recession: first, because there are more defaults than usual; and second, because we will have lower than average recovery rates.

Thus, we should stress the recovery-rate assumptions of our credit risk models when we consider recession scenarios. This is confirmed in Table III.B.2.2, which shows average recovery rates for different phases of the business cycle. In the years 1982–2000 (a proxy for a ‘long-term average’) recoveries tended to be much higher than in the recent (2001 and 2002) downswing, where also the default incidence has increased significantly. For example, recoveries on senior unsecured debt dropped from a long-term average of 43.8% to 35.5% and 34.0% in the 2001/02 recession.

Table III.B.2.2: Average recovery rates (%) of defaulted debt at different periods in time

Asset class	1982–2002	1982–2000	2001	2002
Secured bank loans	61.6	67.3	64.0	51.0
Equipment trust	40.2	65.9	NA	38.2
Senior secured	53.1	52.1	57.5	48.7
Senior unsecured	37.4	43.8	35.5	34.0
Senior subordinated	32.0	34.6	20.5	26.6
Subordinated	30.4	31.9	15.8	24.4
Junior subordinated	23.6	22.5	NA	NA
All bonds	37.2	39.1	34.7	34.3

Source: Moody's KMV (2003)

A popular mathematical model for random recovery rates is the beta distribution. The beta distribution is a distribution for random variables with values in $[0, 1]$ which has the density

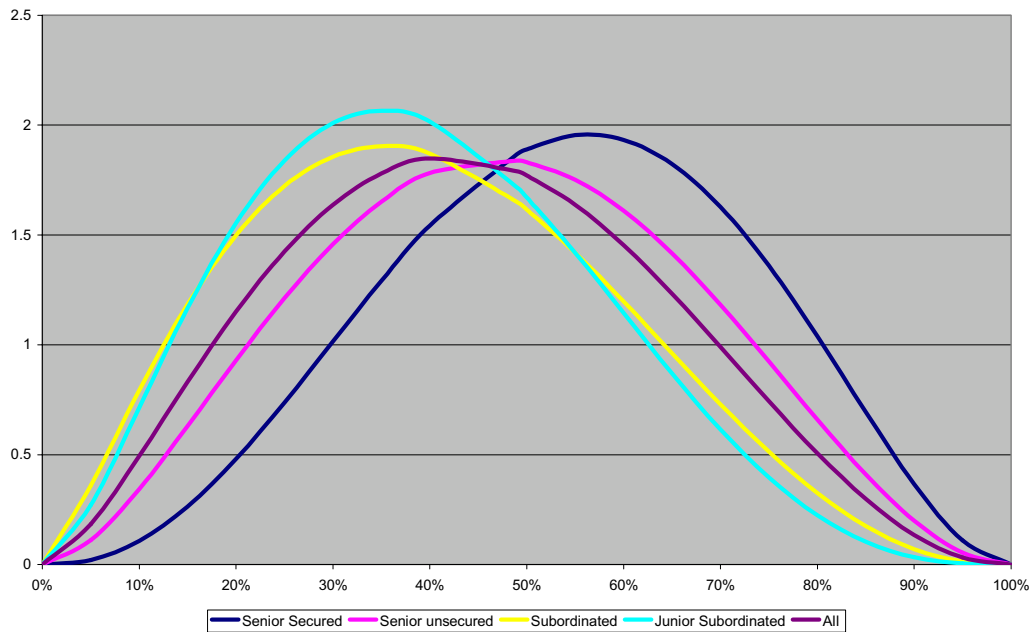
$$f(x) = c \cdot x^a (1-x)^b, \quad (\text{III.B.2.8})$$

where a and b are the two parameters of the distribution, and c is a normalisation constant. By choosing different values for a and b , a large variety of shapes for the recovery distribution can be reached. One can directly fit the parameters of the beta distribution to the mean and the variance of the dataset using the following formulae:

$$a = \frac{\mu^2(1-\mu)}{\sigma^2} \quad \text{and} \quad b = \frac{\mu(1-\mu)^2}{\sigma^2}.$$

Here, μ is the mean of the data, and σ is its standard deviation. In Figure III.B.2.2, the beta distributions fitted to the data in Table III.B.2.1 are plotted. Recovery-rate distributions for senior, senior unsecured, subordinated, junior subordinated, and all debt issues are shown. Quite clearly, higher seniority classes tend to have higher average recoveries, but there is a large overlap between the distributions of the various seniority classes. This is caused by the relatively large variance of the values reported for each class. This can happen even if priority rules of the seniority classes are observed for each obligor individually.

Figure III.B.2.2: Beta distributions fitted to the recovery data of Table III.B.2.1



III.B.2.7 Conclusion

Modelling credit risk has become an essential tool for modern risk management within a financial institution. Such models are used to determine both expected loss (important for pricing loans and other assets or contracts) and unexpected loss (necessary for assessing the appropriate size of the capital buffer).

The fundamental idea presented in this chapter is that probability of default, exposure amount and loss given default (or conversely, recovery rates) combine to give the credit loss distribution for a portfolio of assets. This distribution is typically skewed, with a small probability of large losses.

Recovery rates are one of the three crucial elements in determining the credit loss distribution. They will vary according to seniority, level of security, the economic cycle and local bankruptcy laws, amongst other things. Recovery rates are particularly difficult to estimate in advance, having large standard deviation and dependence on the business cycle.

References

Altman, E I, Resti, A, and Sirone, A (2001) Analyzing and explaining default recovery rates. Report submitted to the ISDA, Stern School of Business, New York University, December.

Gupton, G M, Gates, D, and Carty, L V (2001) Bank-loan loss given default, in *Enterprise Credit Risk Using Mark-to-Future*. Algorithmics Publications, pp. 69–92. Available from www.algorithmics.com

Renault, O, and Scaillet, O (2004) On the way to recovery: A nonparametric bias-free estimation of recovery rate densities. *Journal of Banking and Finance*, **28** (to appear).

Van de Castle, K, Keisman, D, and Yang, R (2001) Suddenly structure mattered: insights into recoveries from defaulted debt, in *Enterprise Credit Risk Using Mark-to-Future*. Algorithmics Publications, pp. 61–68.

III.B.3 Credit Exposure

Philipp Schönbucher¹

III.B.3.1 Introduction

Chapter III.B.2 established the three components of credit loss upon default of an obligor: exposure amount, loss given default (or, conversely, recovery rates) and probability of default. Successful credit risk management and modelling will require an understanding of all three components. Accordingly, this chapter explores in more detail the exposure amount, which may also be referred to as credit exposure or exposure at default. The exposure at time t is the amount that we would lose if the obligor defaulted at time t with zero recovery.

Section III.B.3.2 will distinguish between pre-settlement and settlement risks, as the management techniques vary depending on whether default occurs prior to or at the time of settlement. Section III.B.3.3 explains exposure profiles, that is, how exposures vary over time for the various asset and transaction types. Finally, Section III.B.3.4 discusses some techniques for reducing credit exposures (risk mitigation techniques) and Section III.B.3.5 concludes.

The exposure is closely related to the recovery rate in the sense that it is usually identified with the legal claim amount or the book value of the asset. But this identification is not quite correct: We should also recognise that we might stand to lose value which is not recognised as a legal claim amount (see the discussion in Section III.B.2.6), for example large future coupon payments. Thus, the correct definition of exposure is the market price or the replacement value of the claim.

In practice many simplifications are used when exposure amounts are estimated. This is partly justified by the fact that any accuracy that is gained by a more accurate measurement of exposure will probably be swamped by the large uncertainty surrounding the recovery rates of the obligors. Exposure and recovery enter the loss at default multiplicatively. Thus if we decrease the relative error in the exposure measurement by 5%, it will not help much in improving the loss estimation if on the other hand we face a relative error of 20% or more in the recovery rate.

Credit exposure may arise from several sources:

¹ D-MATH, ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland.

Direct, fixed exposures. These arise from lending to the obligor or from investment in bonds issued by the obligor (which is another kind of lending). This is the most straightforward type of exposure.

Commitments. Although they frequently have zero current exposure, committed lines of credit constitute large potential exposure because they will usually be drawn should the obligor get into financial difficulties. The bank may have covenants that allow a termination of the lending facility in the event of an adverse change of the obligor's credit quality, but the borrower usually has an information advantage and can draw at least parts of the line of credit before his financial problems become known to the bank. This raises the question how the potential exposure embedded in a line of credit should be measured. A common pragmatic solution is to assume that the obligor will have drawn a certain fraction of the line of credit if he should default. Thus, we consider a fixed fraction of a committed line of credit as exposure at default, even if it is not drawn at the moment.

Variable exposures. These arise mainly from over-the-counter (OTC) transactions in derivatives. As they are exposed to (uncertain) moves in the interest rates, fixed-coupon loans and bonds could also be considered variable exposures, but they are usually considered fixed exposures. Futures contracts are generally ignored for the purposes of credit assessment as their institutional features are designed to effectively eliminate credit exposure. The futures clearing mechanism interposes the clearing house as the ultimate counterparty to all trades. The clearing house in turn manages its credit exposure by trading on a fully collateralised basis through the system of initial and daily margin calls (see Section I.C.6.3.1). While it is theoretically possible that a clearing house might default on its obligations as counterparty, it has never actually happened and is generally regarded as a remote possibility.

For OTC transactions, current exposure is defined as the current replacement value of the relevant derivative contract (after taking netting into account). Difficulties arise when it comes to future exposures because this involves the projection of the future value of the derivative conditional on the occurrence of a default while recognising the effects of any netting agreements which may be in place. The future value of the derivative depends, in turn, on market movements which cannot be accurately predicted in advance.

For pragmatic reasons, most credit portfolio risk management models currently map future exposures at default into *loan equivalent exposures*. In this method, the random future exposure of an OTC derivatives transaction is mapped into a non-random exposure (possibly time-

dependent). For credit risk assessment, the derivatives transaction is then treated as a loan with this exposure.

III.B.3.2 Pre-settlement versus Settlement Risk

In Section III.B.3.3 on exposure profiles we shall see that it is not unusual in OTC transactions for very large cash flows to change hands at the settlement of the transaction, even if the net value of the transaction is much smaller than these cash flows. Thus it makes sense to distinguish between the risks that arise specifically at the settlement of these transactions, the settlement risk, and the pre-settlement risk of the transaction before its maturity. The techniques used for managing each of these risks vary.

III.B.3.2.1 Pre-settlement Risk

This is the risk that the counterparty to a transaction (e.g. an OTC derivative transaction) defaults at a date before the maturity (settlement) of the transaction. Here, the existence of early termination clauses is crucial in the reduction of the exposure due to pre-settlement risk. The right of early termination is the credit risk equivalent of ‘cutting your losses’ by closing out a loss-making market exposure. Usually, these clauses are incorporated in the master agreement between the counterparties. They involve triggers due to

- failure to perform on this or a related contract,
- rating downgrade (usually to a class below investment grade),
- bankruptcy.

Should one of these events occur, the contract is terminated and settled immediately, with final payment being the replacement value of the contract (i.e. the value of an otherwise identical contract with a non-defaulted counterparty). If the replacement value of the contract is positive to the defaulted counterparty, it will receive the final payment. Otherwise, the defaulted counterparty will have to make the final payment to the non-defaulted counterparty. In the latter case the defaulted counterparty may default on this final payment, too, which means that the non-defaulted counterparty will have to enter the bankruptcy proceedings with a legal claim amount of the final payment. Thus, the pre-settlement exposure is

$$E(t) = \max\{0; \text{Replacement value at time } t\}. \quad (\text{III.B.3.1})$$

Thus, the exposure for pre-settlement risk is only the replacement value of the derivative, and only if this value is positive to us. The exposure of the settlement risk on the other hand is roughly equivalent to the gross exposure of the transaction.

III.B.3.2.2 Settlement Risk

Many financial transactions between two counterparties involve two simultaneous payments or deliveries – for example, counterparty A delivers a bond and counterparty B delivers the purchase price for this bond (in a straightforward cash-trading transaction), or one counterparty delivers USD and the other counterparty delivers EUR (in a spot FX transaction). Similar final payments are made at the maturity of FX swaps (exchange of principal) or at the maturity of forward contracts.

Settlement risk arises only at the final settlement of a transaction if there are timing differences between the two payments of the transaction. For example, an FX transaction may involve a payment in EUR by bank A, which is made at 9am in London, and a payment in USD by bank B at the same day, which must be made in New York. Because New York is five hours behind London, this payment will be made later than the EUR payment. Should bank B default after receiving bank A's payment, but before making its own payment, bank A will have to try to recover its claim in the bankruptcy court.

A famous example of settlement risk is the case of the German bank Herstatt, which on 26 June 1974 had taken sizeable foreign currency receipts in Europe but went bankrupt (it was shut down for insufficient capital by the German office for banking supervision) at the end of the German business day before it settled its USD payments in New York.

The effect of *settlement risk* is that bank A has a very large exposure (the total notional amount of the transaction) but only for a very short period of time (a few hours). Compared to pre-settlement risk, the difference in exposure size can be very large. This risk can be mitigated by improving the clearing and settlement mechanisms, netting agreements to minimise cash flows (in particular, the cash settlement of price differences instead of physical delivery), or the introduction of a central clearing house which takes both sides' payments in escrow.

Other risk management tools for both settlement and pre-settlement risks are explained in Section III.B.3.4.

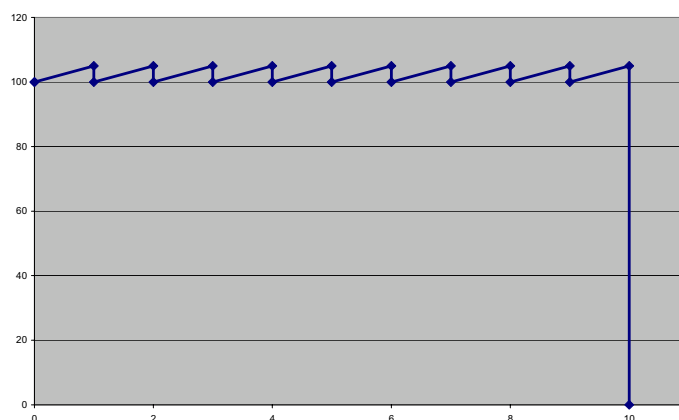
III.B.3.3 Exposure Profiles

III.B.3.3.1 Exposure Profiles of Standard Debt Obligations

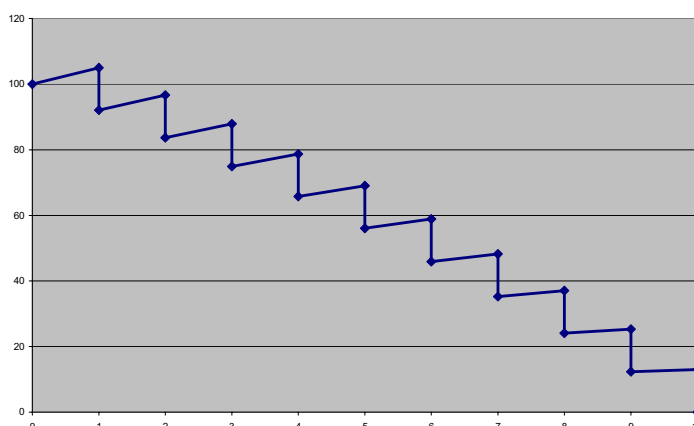
Standard debt contracts usually have fairly straightforward exposure profiles. The simplest case is a bullet bond or a bullet loan with a fixed notional amount paid off at maturity of the loan (see Chapter I.B.2). The top panel of Figure III.B.3.1 shows the exposure profile over time of a ten-year, 5% coupon loan where we have assumed constant interest rates of 5%.

Figure III.B.3.1: Bond exposure profiles

Exposure profile for a ten-year fixed-coupon bond/loan with 100 notional amount, bullet principal repayment at maturity and an annual, fixed coupon of 5%.



Exposure profile for a ten-year amortising loan with 100 notional amount, and annual payments of 12.95. Interest rates are constant at 5% .



Whenever a payment is received, the exposure drops by the payment amount. This causes the characteristic ‘sawtooth’ pattern in Figure III.B.3.1 and also in all other exposure profiles for assets with intermediate payoffs. Between payment dates the exposure increases smoothly. This reflects the increase in the time-value of the outstanding payments.

For the coupon bond in the top part of Figure III.B.3.1, the largest payment is the final repayment of principal (with the final coupon), thus the exposure profile remains largely constant with a large drop in the exposure at maturity. A common approximation is to set the exposure constant until maturity, that is, to ignore the sawtooth pattern caused by the coupon payments and to use some average of the exposure level. Essentially this is equivalent to assuming that defaults only occur in the middle between two coupon payment dates.

Not all loans have the full principal repayment at maturity of the transaction. *Amortising loans*, for example, spread the principal repayments over the life of the transaction in such a way that the total payment amount is the same at all payment dates. This yields the downward sloping exposure profile in the bottom part of Figure III.B.3.1. The exposure profile of amortising loans is slightly concave because, initially, less of the annual payments goes towards principal repayments than at later dates.

III.B.3.3.2 Exposure Profiles of Derivatives

OTC derivative contracts such as swaps, forward contracts or FX transactions have several special features that complicate the calculation of the corresponding exposure amounts. Derivatives often have an initial value of zero (or close to zero). This means that current exposure is a very bad measure of future exposure, which can vary dramatically with market movements. We must consider exposure profiles over time and cannot set exposure to a constant value as we did for fixed coupon bonds.

These exposure profiles are not only time-dependent but also stochastic. Here, assumptions must be made in order to decide how multiple possible realisations of the exposure are to be captured by a single number. At future dates, an OTC derivative can have a positive or a negative value. But by equation (III.B.3.1) the exposure is floored at zero. By definition, exposure cannot become negative. The exposure is the amount that we would lose if the obligor defaulted. So, for instance, with an interest-rate swap, at each payment date either we owe the other party money or the other party owes us money. In the first case our exposure to default of the other party is zero; in the latter case it is positive. Thus we have to cope with an inherent nonlinearity here. The volatility of the underlying asset will also enter the calculation as it defines the range of likely future outcomes for the asset and its derivative.

Derivatives usually involve payments by both contract parties. Thus, it makes a difference whether the two payments are netted (then the exposure is only to the net value of the payments), or whether both payment streams are considered in isolation (in which case we are exposed to the full payment stream of our counterparty).

For the current exposure of a derivative with replacement value $D(t)$ at time t , the starting point for the calculation of derivatives exposures is

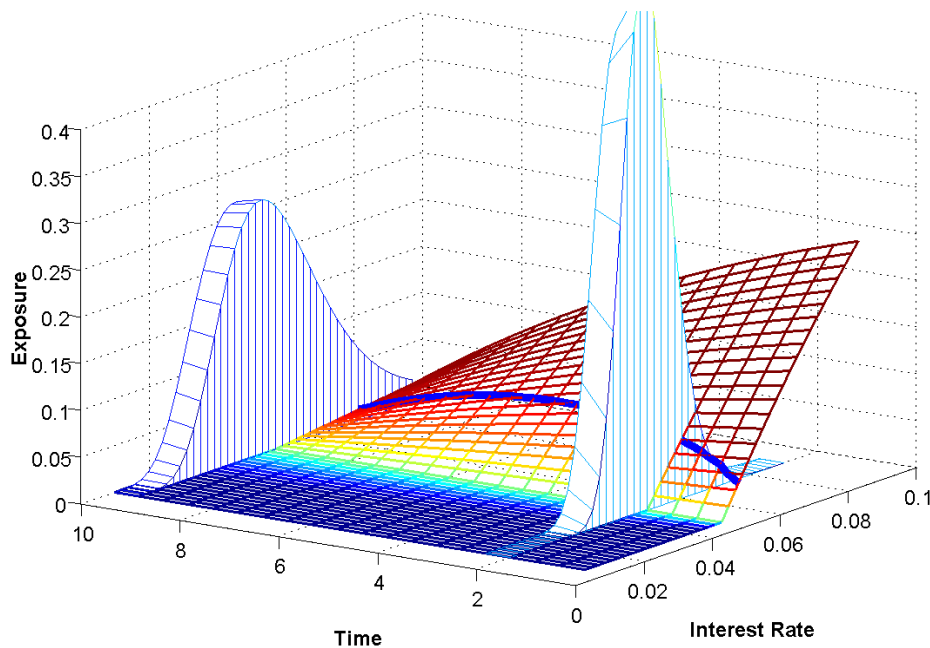
$$E(t) = \max\{0; D(t)\}. \quad (\text{III.B.3.2})$$

If we assume that $D(t)$ can be determined at any time t , there is no inherent difficulty in equation (III.B.3.2) yet.

The difficulties arise when we consider future exposures because exposure is a random variable and, since we usually measure credit risk over a long time horizon (such as one or five years), derivatives exposures may change significantly over the time horizon. Put another way, for a future point in time $T > t$ we cannot predict $D(T)$ with certainty given information at time t , so the exposure will be stochastic.

The exposure measurement problem is now to find a curve $E(t, T)$ which for all relevant $T > t$ maps the distribution of the spot exposure $E(T)$ to a single number: the forward-looking exposure $E(t, T)$ given information at time t . Viewed as a function of time horizon T , the forward-looking exposure $E(t, T)$ is called the *exposure profile* of the transaction.

Figure III.B.3.2 illustrates this using the example of an interest-rate swap with a 5% swap rate from the point of view of the receiver of the floating rate. The surface plot shows the current exposure (i.e. the positive part of the value of the replacement value of the swap) at different points in time and for different possible levels of the interest rates. Unfortunately, seeing this from $t = 0$, we cannot predict future levels of interest rates, we can only make statements over the probability distribution of the interest rates at different time horizons. Figure III.B.3.2 shows the density of the interest rates at $t = 2$ and $t = 10$, and equal colours of the mesh correspond to equal quantiles of the interest-rate distribution. Clearly, uncertainty increases over time, so the density for $t = 10$ is wider than the density for $t = 2$. Because the underlying interest rate is random, the exposure is random, too: it could take any positive value. The bold blue line in Figure III.B.3.2 represents the 90% *quantile exposure*; that is, at any time in the future, the exposure will be below the blue line with 90% probability. The simplification is now to assume that this 90% point will be the realisation of the exposure, in the hope that this assumption – while clearly wrong – will at least be conservative.

Figure III.B.3.2: Interest-rate swap exposure profile*Exposure distribution and exposure profile for a ten-year interest-rate swap at 5%*

Assigning a fixed number to the (stochastic) exposure at time T is a simplification, similar to assigning a single risk measure such as value at risk (VaR) to the full distribution of a random variable. But this simplification allows us to map the complex derivative contract to a *loan-equivalent exposure* amount, so we can view the derivative as a loan with an admittedly rather strange amortisation schedule. Nevertheless, this clever mapping allows us to integrate derivatives contracts into a risk management system which otherwise could only accept loans.

An obvious candidate for an exposure measure is the expected exposure

$$E(t, T) = E_t [(D(t))^+], \quad (\text{III.B.3.3})$$

where $(x)^+ = \max\{x, 0\}$ is the common shorthand notation for the positive part of x and $E_t[\cdot]$ denotes the conditional expectation given information up to time t . Expected exposure has the advantage of being comparatively easy to evaluate. Essentially the problem is reduced to the calculation of the value of a European call option on $D(T)$ – see Section I.B.5.1. In many cases, this calculation can be done quite easily. Expected exposure has the additional advantage of being *coherent* in the sense of Artzner *et al.* (1999). That is, expected exposure is:

- *non-negative*: any derivative with a non-negative replacement value will generate a non-negative exposure number;
- *homogeneous*: a positive scaling of the derivative position will result in the same positive scaling of the exposure measure; and

- *subadditive*: if you add two (or more) derivatives, the joint expected exposure is less than the sum of the individual exposures.

Despite its nice properties, it is frequently felt that the expected exposure is not conservative enough because in many cases the actual exposure at default will be larger than the expected exposure. An alternative, yielding higher exposure values, is to use a ‘quantile-based’ exposure measure. These measures are defined very much like VaR levels: *the p -quantile exposure at time T is the level below which the exposure will remain with probability p* . We denote this by q_p^* , so:

$$P_t(D(T) \leq q_p^*) = p \quad \text{and} \quad E(t, T) = q_p^*. \quad (\text{III.B.3.4})$$

Frequently, the price of the derivative is a monotonic function of the underlying asset and the distribution of the underlying asset is assumed known (e.g. it is lognormal). In this case the p -quantile exposure of the derivative can be calculated by:

- determining the corresponding quantile in the ‘bad’ direction of the underlying asset, and
- calculating the value of the derivative at this level of the underlying asset.

This is usually simpler than calculating the expected exposure at the same level. These calculations have much in common with the calculation of risk measures in market risk management: expected exposure is roughly equivalent to expected shortfall (see Section III.A.3.5.2) and quantile-based exposure measurement is very similar to VaR-based measures of market risk (see Chapter III.A.2). There are some practical difficulties, in that for credit exposure measurement these numbers must be calculated at several time horizons, and for longer time horizons, than the ones commonly used in market risk.

Figure III.B.3.3 shows the exposure profile of an interest-rate swap contract (see Chapter I.B.4). The typical feature of most swap contracts is that they are initially entered at zero exposure: both sides of the swap have the same value. The potential exposure increases as we look further forward in time. This so-called ‘diffusion effect’ is caused by the randomness of the underlying interest rate: the range of possible outcomes increases with time. But as time proceeds, the number and total notional amount of the remaining payments decrease, so that eventually the future exposure has to decrease back to zero. From a statistical point of view we can say that the worst time for the counterparty to default is around seven years into this ten-year swap. The potential for large credit losses at this point exceeds the potential for large credit losses at any earlier time because the range of likely market movements is greater. A large and favourable market movement could mean the loss of a very valuable asset should the counterparty default at

this time. Beyond the seven-year point, the amortisation effect dominates over the diffusion effect, so potential losses are reduced.

Figure III.B.3.3: 95% exposure profile of an interest-rate swap

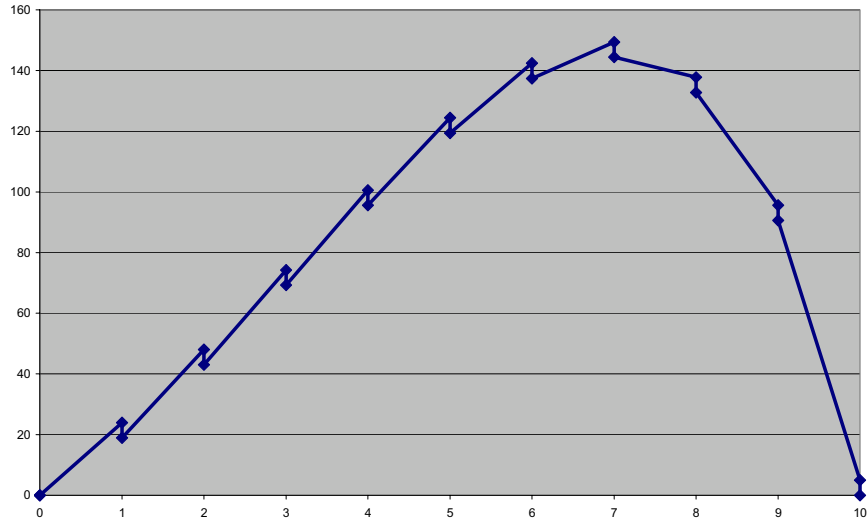


Figure III.B.3.4: 95% exposure profile of an FX swap

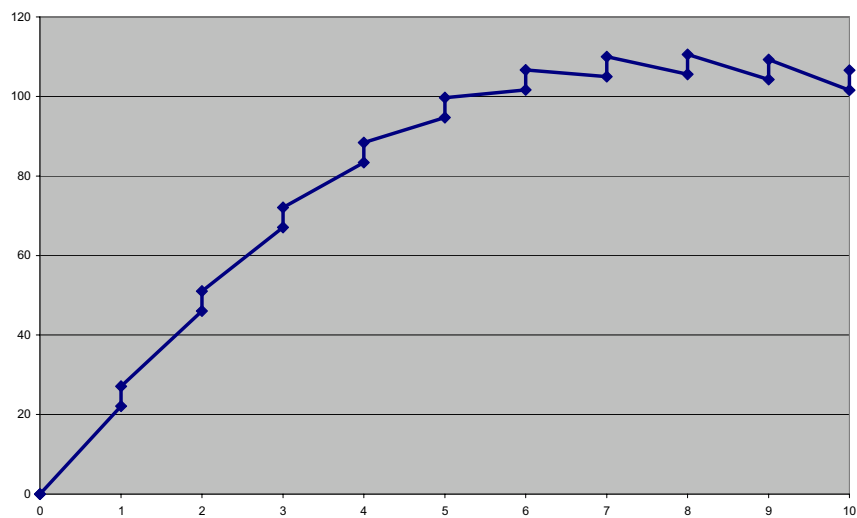


Figure III.B.3.4 shows the typical exposure profile of an FX swap. An FX swap differs from an interest-rate swap in that there are no interim cashflows and therefore no amortisation of risk. The exposure profile reflects only the diffusion effect; the greater the passage of time, the greater the potential for a large favourable exchange rate movement. If the counterparty defaults when the swap is in profit, then a significant asset may be lost. The exposure profile tells us that statistically, the worst possible time for default is towards maturity when the potential mark-to-

market value is greatest. Note also that an FX swap, unlike an interest-rate swap, also has a final exchange of principal. Consequently the settlement risk at maturity is significantly greater.

In Table III.B.3.1 we present a numerical example to illustrate the calculation of the exposure levels of a fixed-for-floating interest-rate swap from the point of view of the floating-rate receiver with a fixed swap rate of 5% and a notional of 100. We assume that the interest rates follow a lognormal random walk with drift zero and volatility 5%, and that the term structure of interest rates remains flat at all times.

Table III.B.3.1(a) shows the calculation of current exposure for one simulated path of the interest rates (shown in the column ‘floating’). In years 1 and 2 interest rates rose above the fixed rate of 5% so that the swap has a positive value to us. The value of the swap is calculated by calculating the value of an annuity that pays a fixed payment of \$1 for the remaining life of the swap, and then multiplying it by the difference between the current interest rate and 5%. (This calculation yields the value of the net payments if we were to enter an offsetting swap.) We see that the value of the swap turns negative from year 3 onwards. Thus, the current exposure is zero in those years, and it is only positive when the swap also has a positive value.

Unfortunately, to carry out the calculations in Table III.B.3.1(a) we need to know the future development of the interest rates, so it cannot be done at time $t = 0$ to generate an exposure profile. The calculations to generate an exposure profile are shown in Table III.B.3.1(b). For the floating-rate receiver, the exposure is worst (highest) whenever interest rates are high. So, in order to calculate a 95% quantile exposure, we need to calculate the upper 95% quantiles of the interest rates; that is, for each year we calculate those levels of interest rates which are *not* exceeded with 95% probability. These levels we can calculate already at time $t = 0$; they are shown in the second column. The rest of the exposure calculation now proceeds as for the calculation of current exposure: we calculate the values of the annuities for these interest rates, and then the value of the swap. As the value of the swap is always positive to us, this is also the exposure in these scenarios. Again, we see the characteristic shape of a hump-shaped exposure profile similar to the one in Figure III.B.3.3.

Table III.B.3.1: Exposure calculations for an interest-rate swap

(a) *Current exposure (simulated interest-rate path)*

Time	Floating	Fixed	Value of annuity	Value of swap	Current exposure
0	5.00%	5%	7.77	0.00	0.00
1	5.30%	5%	7.06	2.12	2.12
2	5.17%	5%	6.47	1.12	1.12
3	4.61%	5%	5.91	-2.31	0.00
4	4.55%	5%	5.19	-2.35	0.00
5	4.74%	5%	4.40	-1.13	0.00
6	4.61%	5%	3.61	-1.40	0.00
7	4.13%	5%	2.79	-2.42	0.00
8	3.99%	5%	1.90	-1.92	0.00
9	3.88%	5%	0.97	-1.09	0.00
10	4.06%	5%	0.00	0.00	0.00

(b) *95% quantile exposure profile*

Time	Floating: upper 95% quantile	Fixed	Value of annuity	Exposure (value of swap at 95% quantile)
0	5.00%	5%	7.77	0.00
1	5.42%	5%	7.03	2.96
2	6.08%	5%	6.24	6.71
3	6.98%	5%	5.44	10.77
4	8.19%	5%	4.64	14.80
5	9.78%	5%	3.86	18.44
6	11.87%	5%	3.09	21.23
7	14.63%	5%	2.34	22.53
8	18.27%	5%	1.60	21.24
9	23.13%	5%	0.84	15.27
10	29.62%	5%	0.00	0.00

III.B.3.4 Mitigation of Exposures

Whenever possible, both counterparties to an OTC derivatives transaction should engage in exposure-minimising agreements. This is a win-win situation for both sides because with well-designed agreements, counterparty exposure (and thus credit risk) can be significantly reduced without needing economic capital and without large additional costs to the counterparties. In this section we consider the problem of aggregating all exposures to a particular counterparty, where netting and other mitigation agreements may be in place.

Note that for the calculation of the total exposure profile with respect to a counterparty we have to consider a whole portfolio of derivatives transactions. This usually means that we will no longer be able to calculate expected exposures with closed-form solutions for European options on the underlying transaction, nor will we be able to identify the ‘critical values’ for quantile-based exposures, as these values will now be a surface in a multidimensional risk space. Thus, the

calculation of exposure profiles is reduced to Monte Carlo simulations or other numerical methods (see Chapter II.G).

III.B.3.4.1 Netting Agreements

To illustrate the magnitude of the benefits of netting agreements, one need only look at the latest OTC derivatives market statistics released by the BIS in Table III.B.3.2. The presence of netting agreements reduces the gross credit exposure of all outstanding OTC derivatives contracts to 28% of the total market value of these contracts. This risk-reduction effect can be even stronger when the two counterparties are involved in a large number of transactions – for example, if they are both market makers in certain OTC markets, or for a very active trader/hedge fund with his prime broker. If the transactions are ‘hedged’ transactions (i.e. delta-hedged derivatives positions) then netting can eliminate almost all exposure.

Table III.B.3.2: The effect of netting on global OTC derivatives exposure (USD bn)

Total outstanding notional amounts	197,177
Total market value of outstanding contracts	6,987
Gross credit exposure after netting	1,986

Source: BIS market statistics, End of December 2003.

A necessary requirement before netting can be applied is the existence of a legally watertight netting agreement, which is usually embedded in a master agreement between the two counterparties.² A *master agreement* is a contract that sets the framework under which the counterparties can undertake derivatives transactions. Each individual transaction is economically independent, but it is governed by the same master agreement. Legally, all transactions are part of this one master contract. Thus, the master agreement can specify rules that apply across several transactions.

The form of netting advocated by ISDA is called *bilateral close-out netting*: Once a credit event occurs on one of the counterparties, all transactions under the netting agreement are closed out. Closing out means that the current market value (replacement value) of all transactions is determined, then these numbers are netted, and then the net amount becomes immediately due.

Suppose counterparty A has entered swaps with counterparty B with current market values (to A) of +34, -97, +45 and +5. Then the net value of these four transactions is -13. If A defaulted, B

² The International Swaps and Derivatives Association (ISDA) was and is engaged in proposing and promoting legislation to enable netting agreements, and such legislation has been adopted in most OECD countries. More legal background information can be found on the ISDA’s web site (www.isda.org); see also Werlen and Flanagan (2002).

would have a claim of 13 against A with which it would go to the bankruptcy court. At an assumed recovery rate of 40%, B would suffer a loss of 7.8. If counterparty B defaulted, then A would pay the net value 13 to B and would not have any further obligations. In this case, A does not have any immediate credit exposure to B, thanks to the existence of the netting agreement.

If, on the other hand, there is no bilateral close-out netting agreement, then B would have claims of 97 against A which B would have to try and recover in bankruptcy court, while B would still have to perform on the other three contracts with a total market value of 84. Thus, B's loss would be significantly larger: at 40% recovery B would lose 58.2. The situation without netting is also bad to A: while A had no net exposure with netting, without netting A does have significant losses. A would have to perform on the -97 swap but would have to try and recover some of the total value of 84 of the other three contracts. The credit losses to A would be 50.4.

The reason for the efficacy of netting is that it prevents cherry-picking by the administrator of the defaulted counterparty. In many legislations, the administrator (receiver/bankruptcy court) can decide only to enforce contracts which are beneficial to the defaulted entity while sending all other contracts and obligations to the bankruptcy courts. The introduction of a master agreement ensures that all swap transactions are viewed as one contract, which can only be accepted or rejected as whole.

From a risk manager's point of view, the existence of a netting agreement with a certain counterparty allows him to aggregate all exposures with that counterparty and to consider only the net exposure that arises from these transactions. He is able to calculate an exposure profile with respect to the counterparty and not just with respect to an individual transaction.

III.B.3.4.2 Collateral

Apart from netting agreements, collateral is another popular instrument to mitigate counterparty risk, in particular if one counterparty (e.g. a hedge fund) has a significantly more likely to default than the other (e.g. an investment bank). The International Swaps and Derivatives Association carries out regular surveys on the use of collateral and estimates that at the beginning of 2004, around USD 1017bn of collateral were used in OTC derivatives transactions. Comparing this to the USD 1986bn of credit exposure after netting found by the BIS, we see that about half of the non-netted credit exposure is managed by collateral. Another, less obvious form of collateralisation arises when derivatives are embedded into bonds or notes, as for example in credit-linked notes or equity-linked notes. By buying the note, the investor implicitly posts full collateral for all possible exposures that may arise from the derivative which is embedded in the

note. This explains why this is a very popular structure for derivatives transactions with retail customers.

Like netting, collateral also requires a corresponding legal document (again the ISDA publishes template contracts on this topic), but here the legal difficulties are less severe as collateral is a very old means of mitigating credit risk.

Generally, as in netting agreements, the idea is to reduce credit risk by reducing exposure. In the case of collateral, the credit risk of the counterparty is enhanced by the credit risk of the collateral. The lender only suffers a loss if both counterparty and collateral default.

Let us assume that hedge fund A and investment bank B want to enter an OTC derivatives transaction such as an interest-rate swap. To mitigate A's credit risk, the parties enter a collateral agreement which specifies which assets may be delivered by A as collateral, and under which conditions B may ask for collateral, including the amount of the collateral. Typical collateral assets are cash (used in about 70% of cases) and government securities (about 15% of cases), but other assets are possible (e.g. equities). As the collateral may decrease in market value just when the counterparty defaults, non-cash collateral is not counted at its full face value but at less; that is, it is reduced by a 'haircut' factor depending on the volatility of the collateral and its correlation with the underlying exposure.

Over the life of the swap, A will have to deliver the required amount of collateral assets to B. The assets are legally still the property of A but they are under administration by B. If the collateral becomes insufficient (e.g. because of market movements or because of changes in B's credit rating), B issues a margin call asking A to post additional collateral. If there is excess collateral, A may remove the collateral from the collateral account. If at the end of the transaction A has not defaulted, all unused collateral is returned to A.

If on the other hand counterparty A misses a payment of the swap at some point, then B is entitled to sell some of the collateral assets to make the payment to himself. If an early termination event occurs because of a default of A, then B is allowed to sell the full collateral in the market and use the proceeds to cover the replacement value of the contract. If there are any remaining proceeds from the sale of the collateral, these are returned to A. If the collateral was not sufficient to cover the replacement value of the derivative, the remaining value will have to be recovered in the usual way.

For exposure measurement, the value of the collateral must be deducted from the replacement value at risk at every time horizon. If the value of the collateral itself is random this introduces an additional level of complexity, but for the most common case of cash collateral we can simply subtract the collateral amount from the exposure value.

Collateral management is the non-trivial task of keeping track of both the collateral a business has to post, and the collateral it should receive. The institution must ensure that it can always provide sufficient collateral to support its transactions, even if adverse market movements or a credit downgrade suddenly increase the collateral requirements. Thus, a collateral manager also has to monitor the credit quality of his own institution very closely. There are many famous examples of the failure of collateral management policy, including the downfall of the Long-Term Capital Management hedge fund.

III.B.3.4.3 Other Counterparty Risk Mitigation Instruments

Limits. Counterparty exposure limits are not a mitigation instrument. Exposure limits are counterparty risk management devices that are used to avoid undue counterparty risk concentrations with respect to any particular counterparty. They are also used to avoid exposure to potential counterparties that are deemed to be insufficiently creditworthy; in such cases a limit is not granted, preventing any trades or lending activity which might result in credit exposure. In most financial institutions the credit committee is responsible for determining whether the institution is willing to expose itself to performance risk for any particular counterparty/borrower, and to what extent. The existence/size of a limit will either be determined on the basis of its own analysis, or can effectively be outsourced by relying on ratings from ratings agencies. Non-financial corporations also establish counterparty limits, but are more likely to rely on ratings agencies for credit assessments. Separate limits should apply for settlement and pre-settlement risk.

Termination rights/credit puts. One or both counterparties may reserve the right to terminate the transaction should the credit risk of the other party worsen significantly. This allows the recovery of the exposure before the actual default event at a higher (often full) recovery value.

Establishing third-party guarantees. A traditional and popular way to mitigate credit risk is to establish a guarantee for the exposure from a third party. This is very similar to collateralisation but its effect is to replace the default probability (as opposed to the exposure) of the counterparty with the combined default probability of the counterparty and the guarantor.

Credit derivatives. Counterparty exposures can also be managed with credit derivatives (see Chapter I.B.6). Although it is technically possible to specify a credit derivative which pays off exactly the exposure at default of a specific counterparty, this is usually not done in practice for two reasons: first, this would entail a disclosure of the transactions that have been made with the counterparty; and second, because of the unusual nature of the payoff, the credit protection would be rather expensive. Nevertheless if one is willing to accept some residual exposure to the counterparty, a significant part of the counterparty exposure can be laid off using single-name credit default swap contracts.

All counterparty risk mitigation instruments necessarily introduce *legal risk* and *documentation risk*. Collateral calls may not be made or fulfilled in time, and netting agreements and guarantees may be legally challenged. These risks are hard to quantify and must be carefully monitored. Furthermore, these risk mitigation instruments also have a cost in terms of administration costs, which must be justified by the reduction in risk. More details on credit risk management instruments can be found in Chapter III.B.6.

References

- Artzner, P, Delbaen, F, Eber, J-M, and Heath, D (1999) ‘Coherent measures of risk’, *Mathematical Finance*, 9(3), pp. 203–228.
- Werlen, T J, and Flanagan, SM (2002) ‘The 2002 Model Netting Act: A solution for insolvency uncertainty’, *Butterworths Journal of International Banking and Finance Law*, April, pp. 154–164.

III.B.4 Default and Credit Migration

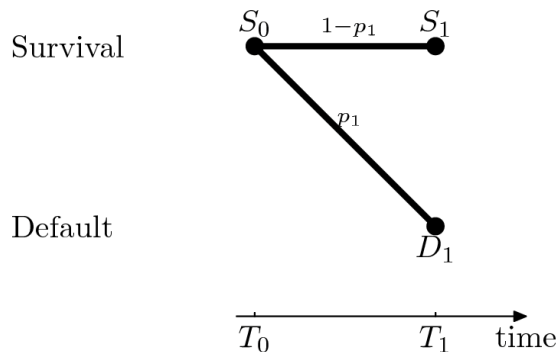
Philipp J. Schönbucher¹

Having covered recovery rates and exposures in the previous chapters, this chapter discusses the modelling and measurement of the third determinant of default loss: the *probabilities* of default of individual companies (and, by extension, sovereign nations). After setting up a framework to describe and analyse default probabilities we will consider three different credit risk assessment methods: agency ratings, internal ratings and market-implied default probabilities. Finally, we compare the three approaches and discuss their differences.

III.B.4.1 Default Probabilities and Term Structures of Default Rates

In this section we introduce the terminology for describing default probabilities. The starting point of any representation of default probabilities is the one-period default probability depicted in Figure III.B.4.1.

Figure III.B.4.1: A one-step default tree



Starting from node S_0 at time T_0 , there are two possible outcomes at time T_1 in Figure III.B.4.1: default (node D_1) which is reached with the default probability p_1 ; and survival (node S_1) which is reached with the *survival probability*, $1 - p_1$. Apart from the direct specification of the default probability p_1 (or the survival probability $1 - p_1$), we could also specify the *odds of default*. The ‘odds’ of an event are defined as the ratio of the probability of the event (i.e. the default probability) to the probability that the event does not occur (i.e. the survival probability):

¹ D-Math, ETH Zurich, Switzerland.

$$H_1 := \frac{p_1}{1 - p_1}. \tag{III.B.4.1}$$

Hence, given the odds of an event, one can recover the probability of the event as:

$$p_1 = \frac{H_1}{1 + H_1}.$$

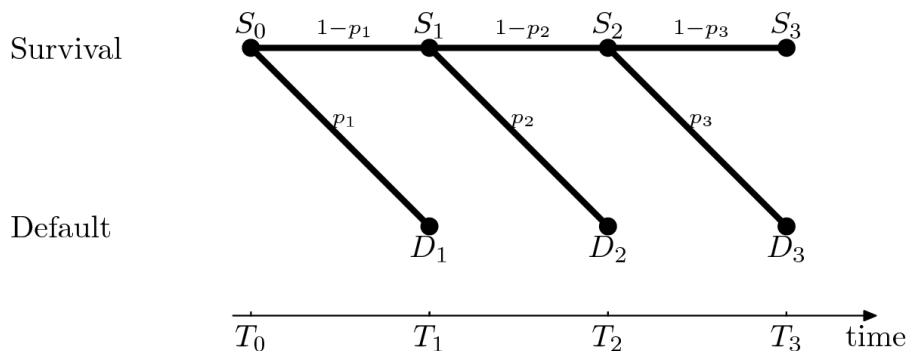
This definition differs only slightly from the ‘odds’ that are routinely quoted by bookmakers on sports (and other) events. In fact, bookmakers usually quote $1/H_1$ and not H_1 . We can interpret H_1 to be the odds of a fair bet on the event, in that if you bet \$1 on the event that the obligor defaults then:

- you lose your \$1 if the obligor survives, or
- you get $1/H_1$ if the obligor does indeed default.

The expected payoff of this bet is $-1 \times (1 - p_1) + p_1/H_1 = 0$, so the bet is fair.

From a modelling point of view the advantage of using odds is mostly of a technical nature: probabilities are restricted to lie between 0 and 1, while odds can take any value between 0 and infinity. For small values (such as for short-horizon default probabilities), odds and probabilities are almost equal, with odds being slightly larger. For example, if the default probability is $p_1 = 2\%$, then the ‘odds’ of default are $H_1 = 2.0408\%$.

Figure III.B.4.2: A schematic representation of default and survival over time



A complication in the representation of default probabilities arises when several points in time are considered. Figure III.B.4.2 shows a simple representation of default and survival over time in a model with three periods ($[T_0, T_1]$, $[T_1, T_2]$ and $[T_2, T_3]$). Over each period $[T_{i-1}, T_i]$, the obligor can default (with probability p_i) or survive (with probability $1 - p_i$).

By considering the individual periods in isolation, we can define ‘local’ default probabilities (and local odds of default) as before. So, we can still analyse the situation in period $[T_1, T_2]$ with default probability p_2 and odds of default $H_2 = p_2/(1 - p_2)$. But these quantities are now *conditional* default probabilities, that is, they are conditional upon survival until T_1 . They are only valid if we have reached the node S_1 . If a default has already occurred in the first period, then p_2 and H_2 have no meaning. These conditional default probabilities are also often called *marginal default probabilities*.

If we now want to know the *cumulative default probability* until T_2 , that is, the probability of default at *any* point in time over $[T_0, T_2]$, then we must take several possible paths across the tree into account, which all can lead to a default during the period $[T_0, T_2]$:

- the obligor survives the first period ($S_0 \rightarrow S_1$) and defaults in the second period $S_1 \rightarrow D_2$,
and
- the obligor defaults in the first period ($S_0 \rightarrow D_1$).

The first scenario has probability $(1 - p_1)p_2$ and the second scenario has probability p_1 , so that the total default probability over $[T_0, T_2]$ equals $(1 - p_1)p_2 + p_1 = p_2 + p_1 - p_1p_2$.

If we project further into the future it quickly becomes more convenient to consider cumulative *survival* probabilities because for survival we only have to consider *one* path across the tree. The cumulative survival probability over $[T_0, T_3]$, for example, is given by the product of the marginal survival probabilities over the individual periods which are spanned by the interval $[T_0, T_3]$, that is,

$$P[\text{Survival over } [T_0, T_3]] = (1 - p_1)(1 - p_2)(1 - p_3) = P(T_0, T_3). \quad (\text{III.B.4.2})$$

Clearly, the cumulative default probability over $[T_0, T_3]$ equals one minus the cumulative survival probability over $[T_0, T_3]$. The general formula for the probability of $S_i \rightarrow S_j$, that is, of going from survival in T_i to survival in T_j with $i < j$, is

$$P(T_i, T_j) = (1 - p_{i+1})(1 - p_{i+2}) \cdots (1 - p_j).$$

Starting from the tree in Figure III.B.4.2, we can specify default probabilities (or odds of default) for future time periods, that is, we can specify a *term structure* of default probabilities. In many cases it may be desirable to increase the resolution of the term structure by inserting additional points in time, for example, going from yearly to quarterly, monthly or even daily intervals as this

will allow us to place the nodes of our tree exactly on the payment dates of the bonds or loans of the obligor under consideration.

As the length of the time periods gets smaller, local default probabilities and odds of default should also decrease. For example, suppose we have an obligor with a 10% default probability over one year. To be consistent with this when moving to monthly time periods, we should assume that the monthly default probability is 0.874%. The constraint becomes stronger as we take smaller steps. For instance, when calculating the daily default probability that is equivalent to a 10% one-year default probability, even a small overestimation of the daily probability can lead to a one-year default probability that is much too large. In summary, to avoid strange limiting behaviour as the time period decreases we must reduce the term default probabilities in a coordinated way.

A common assumption that always achieves a stable balance when time-steps are reduced is that the odds of default over a small time interval $[T_i, T_{i+1}]$ are approximately proportional to the *length* of the time interval. Let us denote the length of the time interval by $\Delta = T_{i+1} - T_i$. Then the hypothesis may be written:

$$H_i = \Delta \times \lambda(T_i) \quad \text{or} \quad \lambda(T_i) = \frac{H_i}{\Delta}, \quad (\text{III.B.4.3})$$

where the proportionality factor is denoted by $\lambda(T_i)$. Now, if we take the limit as the time interval gets smaller and smaller (i.e., as $\Delta \rightarrow 0$), we leave $\lambda(T_i)$ unchanged. But the odds H_i of default get smaller, too. $\lambda(T)$ is the default probability per unit of time, evaluated at the gridpoint which corresponds to time T . This quantity is known as the *default rate*, the *default intensity* or the *default hazard rate* at time T ; it is the ‘instantaneous’ probability of default in a continuous time setting.

Suppose that, somehow, we are able to specify a default hazard rate function $\lambda(t)$ for all $t \geq 0$. Armed with this function we can calculate survival and default probabilities from 0 until a given time horizon T . If we then let the interval length Δ go to zero, it can be shown that eventually the survival probability will be

$$P(0, T) = \exp\left\{-\int_0^T \lambda(t) dt\right\}, \quad (\text{III.B.4.4})$$

and the probability of a default between time 0 and time T will thus be $1 - P(0, T)$.

Conversely, suppose we are given a term structure of survival probabilities, that is to say, a set of probabilities $P(0, T)$ for different time horizons T . Then we could compute the corresponding default intensity function by differentiation. In fact:

$$\lambda(T) = \frac{\partial}{\partial T} \ln P(0, T). \quad (\text{III.B.4.5})$$

So, there is a correspondence between the term structure of survival probabilities (or default probabilities) and the default hazard rate function. We may, for instance, try to estimate a term structure of default probabilities using one of the methods described later in the chapter. Whatever the shape of this term structure – and market-implied term structures can have quite strange shapes – it is usually possible to find a default hazard rate function that reproduces this shape.²

An important special case arises when the default hazard rate is *constant*, an assumption that is at the foundations of the CreditRisk+ model (see Chapter III.B.5). In this case, let us write the hazard rate as λ_0 . Then by equation (III.B.4.4) the survival probabilities are just:

$$P(t) = \exp\{-\lambda_0 t\}. \quad (\text{III.B.4.6})$$

This gives a simple and effective way to interpolate default probabilities between different dates.

For example, suppose we choose $\lambda_0 = 10.53\%$. Since by (III.B.4.6), $\lambda_0 = -\ln(P(1))$, the one-year survival probability is $P(1) = \exp\{-\lambda_0 T\} = 90\%$ so the one-year default probability is 10%. Also, assumption (III.B.4.3) gives the six-month default probability as 5.13%. Similarly, over one month the default probability will be 0.87% and over one day it will be 0.029%.

III.B.4.2 Credit Ratings

The goal of any credit rating system is the accurate assessment of the credit risk of the obligor. Credit ratings are one of the most important tools to assess the likelihood that an obligor defaults and their use is actively encouraged by the new capital adequacy rules for credit risk proposed by the Basel Committee on Banking Supervision. The aim of a credit rating procedure is the accurate classification of obligors according to their credit quality, usually by specifying an estimate of the obligor's default probability or by giving them a 'letter rating'.

²The only exception are term structures of survival probabilities which have jumps or which drop to zero.

Large public rating agencies such as Standard and Poor's, Moody's and Fitch classify obligors into a number of *rating classes*.³ For example, Standard and Poor's use the classes AAA, AA, A, BBB, BB, B, CCC, D, with the interpretation that 'lower' classes (such as CCC) carry a higher risk of default over a given time horizon than higher classes (AAA or AA).

A classification into a finite set of classes is not necessarily the only output of a credit rating system. Many other systems directly produce estimates of the default probability of an obligor which can vary on a continuous scale, essentially from 0% to 100%. Such systems are typically based upon a quantitative and statistical model of the default risk of the obligor. Examples are KMV's 'expected default frequencies' (see Chapter III.B.5) or Kamakura's default probabilities; most proprietary 'internal' ratings models also fall into this class.

III.B.4.2.1 Measuring Rating Accuracy

There are many different approaches to the problem of determining a credit rating, and it is important to have a methodology to judge the accuracy of the output of the rating procedure. Unfortunately, a credit rating can be correct but still unlucky (it correctly classifies obligors, but some 'good ones' default and the 'bad ones' do not). Conversely, it can be incorrect, but lucky. Distinguishing between these two possibilities can be difficult.

When classifying rating systems one might be tempted to say that the best rating system is the one that produces the 'true' default probability. Unfortunately, the concept of the 'true' default probability is very elusive: it depends on the information that is available, and it has no direct relation to observable quantities.

Imagine we had access to an ideal rating system, let us call it 'Crystal Ball'. Crystal Ball only needs two classes to accurately classify all obligors: D and S. An obligor is classified as D if it will default, and as S if it will survive, and our Crystal Ball can do this without error. Having Crystal Ball, we no longer need default probabilities. There are only two trivial values for default probabilities: 0% (for the S class) and 100% (for the D class). Essentially, these are the only 'true' default probabilities. In reality we do not have a Crystal Ball, so we specify default probabilities

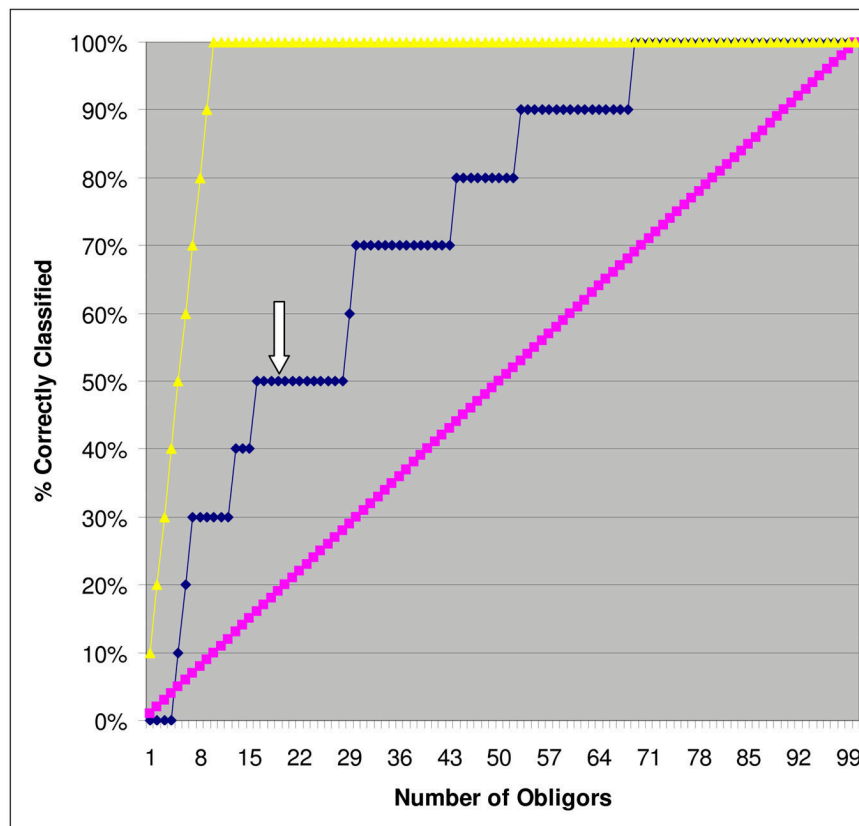
³ Moody's bought KMV and formed Moody's KMV in 2003. Today, Moody's KMV provides both classical Moody's 'letter' ratings and the 'expected default frequency' (EDF) ratings of KMV. As these two rating systems are fundamentally different, we will refer to Moody's ratings when the letter ratings of Moody's KMV are meant, and we will refer to KMV ratings for the EDF equity-based default risk measures calculated according to the KMV methodology.

and/or intermediate rating classes to measure the degree of our confidence about the fact that the obligor truly belongs to class S or class D.

A common method to compare different rating models is to focus on their ability to accurately *rank* obligors according to their credit quality. This method ignores any possible concrete values that are given for the default probabilities, which can be an advantage if the system does not directly give us such probabilities but only qualitative rankings like the letter ratings of the public rating agencies.

Figure III.B.4.3 shows a *cumulative accuracy plot* (CAP) with which the accuracy of the risk ranking of rating models can be compared. The plot is constructed as follows. First, the obligors are ranked according to the default risk that the model assigns to them, starting with those with the highest default risk. The rank numbers will form the x -axis of the plot. The CAP now shows, for each number/percentile x , what percentage y % of the *actually* defaulted obligors can be found among the x % worst obligors according to the model's ranking.

Figure III.B.4.3: A cumulative accuracy plot



For example, the marked point of the blue CAP is found as follows. The basic data set contained 100 obligors, of which 10 defaulted during the next year. We now use our credit model to classify these 100 obligors based upon the data available at the beginning of the year and rank the obligors in order of decreasing credit risk (according to the model's prediction). Next to that, we also keep track of the actual default/survival behaviour of the obligors. The marked value in the graph is at an x -value of 19, which means that we must consider the 19 'worst' obligors according to the model. Out of these 19 obligors, the 5th, 6th, 7th, 13th and 16th actually did default, so that we have captured 50% (5 out of 10) of the actual defaulters in these 19 worst obligors. Thus, the y -value at $x = 19$ is $y = 50\%$. The same calculation is done for every x between 0 and 100. Clearly, any CAP plot must start at 0% (an empty set of zero obligors cannot contain any defaulters) and it must end at 100% (if you have all obligors, you also have all defaulters).

The CAP of the Crystal Ball model is the yellow line, and shows the best possible default prediction accuracy. The ten worst obligors of the Crystal Ball model are also the ten obligors that default in reality. Thus, the accuracy profile shows a very steep increase up to 100% explained (correctly classified) obligors already at the 10th ranked obligor. Beyond that, the CAP for the Crystal Ball model remains flat because there is nothing more to explain.

The other extreme is a 'random' model in which the obligors are ranked completely at random, without any reference to their actual credit quality. With such an approach, we can expect to find 50% of the actual defaulters in the 50% (randomly chosen) 'worst' obligors, 10% of the actual defaulters in the 10% 'worst', etc., because any *randomly* chosen subset of x obligors will on average contain $10x/100$ defaulters. Thus, the proportion of actual defaulters should be proportional to the proportion of the number of obligors selected, and this purely random credit risk ranking produces the diagonal line shown in pink in Figure III.B.4.3.

A good credit risk model will exhibit a CAP profile that is as close to the yellow line of the perfect model and as far away from the pink line of the purely random model as possible. It will concentrate the defaulters at the high risk scores at the beginning of the ranking, so that the slope of its CAP will be steep initially. The closer the CAP is to the yellow line, the better it is. The dark blue line in Figure III.B.4.3 shows the cumulative accuracy profile of an average default prediction model. If we compare two different rating models and the CAP of one model is consistently above the CAP of the other, then this is a sign of superior classification ability for the first model and it will be considered the better model.

A CAP gives a very good visual impression of the predictive performance of a credit risk model. Nevertheless, it is often useful to reduce the CAP to a single number, the *accuracy ratio* (AR). The

AR is the ratio of the area between the CAP of the credit risk model under consideration and the random model's CAP (the diagonal), and the area between the perfect prediction model's CAP (Crystal Ball) and the diagonal. (Here the x -axis is now also a percentage scale.)

The area between the ideal model and the random model is easily found to be $(1 - D)/2$, where D is the fraction of defaulted obligors. Thus the accuracy ratio of any given model with CAP function $CAP(x)$ is:

$$AR = \frac{\int_0^1 CAP(x)dx - \frac{1}{2}}{\frac{1}{2}(1 - D)}. \quad (\text{III.B.4.7})$$

An accuracy ratio close to 1 means that the model is almost as good as the 'ideal' model, while an accuracy ratio of 0 means that it is as bad as a purely random classification. A negative accuracy ratio is possible: it means that the model does an even worse job than a purely random credit ranking and that the model's ranking should be inverted.

The CAP and the AR only measure the model's ability to *rank* obligors, they say nothing about the model's ability to give correct *values* for the probability of default. This is an advantage when we want to compare models that do not necessarily give us numerical probabilities of default at all (like the 'letter' ratings of public rating agencies) or if we want to use the model as a support tool for 'yes/no' loan decisions.

Yet, theoretically, a model which gives completely absurd values for the default probabilities may still have an almost perfect-looking CAP. As an extreme example, consider a model that assigns a default probability of 50.1% to the first ten obligors (the ones who do default in the end) and a default probability of 49.9% to the other 90 obligors. This model correctly ranks the obligors, thus it has a perfect CAP, but the assignment of almost 50% default probability to all obligors is strongly invalidated by the actual experience of only 10 defaults: 10 or fewer defaults out of 100 obligors would have a probability of essentially zero ($\approx 10^{-17}$) if the individual default probabilities were indeed at 50%.⁴ Despite this extreme example, in practical applications we may expect a model that does a good job in ranking the obligors also to provide good estimates for the default probabilities. We simply have to check that aspect of the model also.

⁴This is the value if defaults are independent. But even with reasonable default correlation, we will be able to strongly reject the hypothesis that the default probabilities equal 50%.

III.B.4.3 Agency Ratings

III.B.4.3.1 Methodology

The core business of public rating agencies such as Standard and Poor's, Moody's and Fitch is the analysis of issuers of debt instruments, in terms of their credit quality and their ability to pay their investors. This analysis is summarised in a rating classification into one of several rating classes: Aaa to C for Moody's, and AAA to C for Standard and Poor's and Fitch. Agency rating classifications are publicly available and are an important factor driving the decisions of many potential investors in these bonds.

Issuing a bond (instead of a loan) has the advantage that the issuer is able to borrow many small amounts from a large number of investors simultaneously. These investors will also be able to trade the bond on a secondary market. Unfortunately, a thorough analysis of an issuer's credit quality is usually too expensive compared to the small investment amount of most individual investors. This might prevent them from investing in the bond in the first place, and the issuance of the bond may fail. The secondary market for the bond will also suffer from the same problem. But if the credit research is carried out by an independent and trustworthy agency which acts for all bond investors and which makes the findings of the research public, a duplication of research effort is avoided. Investors can invest many small amounts in many different bond issues and thus diversify their risk without spending unfeasible amounts on research. Hence a recognised public credit rating is almost indispensable for issuance in modern bond markets. Indeed, the cost of the rating is usually paid for by the issuer of the bond.

Because of their crucial position as one of the most important sources of information to investors, rating agencies often gain access to information that may be unavailable to ordinary investors. This information includes such things as direct interviews with the issuing firm's management, internal planning, research and budgeting numbers, and the accounting data of the rated firm. These data are summarised by a credit analyst who then – possibly with the aid of proprietary statistical models – assigns a rating to the issuer and the bond issue. Naturally, rating agencies are rather secretive about the precise methodology used and the weightings of the factors that influence a rating. There is a portion of human judgement involved, but a surprisingly large fraction of the credit ratings can be reproduced using purely statistical models.

Public agency ratings are revised and updated at regular intervals to reflect changes in the credit quality of the rated obligor. Here, rating agencies try to strike a balance between rating stability and rating accuracy. As the rating is meant to represent a long-term view, the effects of general business cycle variations should average out. This 'rating through the cycle' policy is not

uncontroversial because it can lead to delays in rating adjustments. It will also cause empirical difficulties when we try to back out estimates for default probabilities from agency ratings.

The Standard and Poor's ratings from AA to CCC shown in Table III.B.4.1 may be modified by the addition of a plus or minus sign to show relative standing within the major rating categories. Obligors rated BB, B, CCC, and CC are regarded as *speculative-grade* investments, obligors rated BBB and better are regarded as *investment-grade*.

Besides the classical, 'fundamental' ratings, quantitative credit rating providers have more recently entered the market using proprietary quantitative statistical models to produce an output which can be more directly interpreted as a 'probability of default'. Examples are KMV's *expected default frequencies* and Kamakura's default probability estimates. These ratings typically are 'point-in-time' ratings that do not attempt to smoothen business cycle effects.

Table III.B.4.1 Standard and Poor’s long-term issuer credit ratings definitions

- An obligor rated AAA has *extremely strong* capacity to meet its financial commitments. AAA is the highest issuer credit rating assigned by Standard and Poor’s.
- An obligor rated AA has *very strong* capacity to meet its financial commitments. It differs from the highest-rated obligors only in small degree.
- An obligor rated A has *strong* capacity to meet its financial commitments but is somewhat more susceptible to the adverse effects of changes in circumstances and economic conditions than obligors in higher-rated categories.
- An obligor rated BBB has *adequate* capacity to meet its financial commitments. However, adverse economic conditions or changing circumstances are more likely to lead to a weakened capacity of the obligor to meet its financial commitments.
- An obligor rated BB is *less vulnerable* in the near term than other lower-rated obligors. However, it faces major ongoing uncertainties and exposure to adverse business, financial, or economic conditions which could lead to the obligor’s inadequate capacity to meet its financial commitments.
- An obligor rated B is *more vulnerable* than the obligors rated BB, but the obligor currently has the capacity to meet its financial commitments. Adverse business, financial, or economic conditions will likely impair the obligor’s capacity or willingness to meet its financial commitments.
- An obligor rated CCC is *currently vulnerable*, and is dependent upon favourable business, financial, and economic conditions to meet its financial commitments.
- An obligor rated CC is *currently highly vulnerable*.

Source: Standard & Poor’s.

III.B.4.3.2 Transition Matrices, Default Probabilities and Credit Migration

We have seen that public rating agencies only use ‘letter’ ratings in order to classify obligors according to their credit quality. This may be adequate to make relative comparisons and maybe intuitive buy/hold/sell investment decisions, but in order to be able to use these ratings in a quantitative risk management system we need to map the letter ratings to numbers, to default probabilities. Essentially, we face the problem of backing out what the rating actually *means* in terms of default probability; the verbal definitions as they are given in Table III.B.4.1 are not sufficient for this.

**Table III.B.4.2: Standard and Poor’s one-year average rating transition frequencies,⁵
1981–1991 (percentages)**

	AAA	AA	A	BBB	BB	B	CCC	D
AAA	89.10	9.63	0.78	0.19	0.30	0	0	0
AA	0.86	90.10	7.47	0.99	0.29	0.29	0	0
A	0.09	2.91	88.94	6.49	1.01	0.45	0	0.09
BBB	0.06	0.43	6.56	84.27	6.44	1.60	0.18	0.45
BB	0.04	0.22	0.79	7.19	77.64	10.43	1.27	2.41
B	0	0.19	0.31	0.66	5.17	82.46	4.35	6.85
CCC	0	0	1.16	1.16	2.03	7.54	64.93	23.19
D	0	0	0	0	0	0	0	100

Rating agencies may be secretive about their methodology, but fortunately they publish a lot of historical data about both rating transitions and defaults of rated obligors. A typical summary of such data published by Standard & Poor’s is the *transition matrix* shown in Table III.B.4.2; similar transition matrices are also regularly published by other agencies (see, for example, Hamilton *et al.*, 2002). In Table III.B.4.2 we have eliminated the ‘not rated’ class and added the D rating class for defaulted obligors in the last row, using the assumption that a defaulted obligor remains in the default class with 100% probability.

Each row of a transition matrix gives the historical rating transition frequencies for the obligors of the corresponding rating class. For example, according to the BBB row of Table III.B.4.2, on average 0.06% of all BBB-rated obligors were upgraded to AAA in the course of one year, 0.43% were upgraded to AA, 6.56% were upgraded to A, 84.27% did not change their rating, 6.44% were downgraded to BB, 1.60% were downgraded to B, 0.18% were downgraded to CCC, and 0.45% defaulted.

A first glance at Table III.B.4.2 already confirms some stylised facts about rating transitions. First, default frequencies decrease with increasing rating classes, thus the ratings do indeed reflect an ordering according to default probability. Second, the frequencies of unchanged ratings (i.e. the values on the diagonal) are very high; this is an indicator of the ‘rating stability’ aimed for by the agencies.

⁵ The ‘no rating’ category has been eliminated.

As they stand, the numbers given in Table III.B.4.2 are *not* probabilities, they are the *realised frequencies* of the transitions over the observation period. Thus, a zero entry (as in the AA→D cell) does not mean that AA-rated obligors have a zero probability of default, it only means that no AA-rated obligor defaulted in the years 1981–1991. Generally, small transition frequencies are usually based upon a very small number of observations, so that we have to expect inconsistencies like zero probabilities. Also, realised frequencies can be non-monotonic. For instance, in Table III.B.4.2 the CCC-rated obligors had a higher upgrade frequency to A than BB-rated obligors: 1.16% for C-rated obligors as opposed to only 0.79% for B-rated obligors! Clearly, transition frequencies can only be considered as noisy estimates of the true transition probabilities.

In order to use the information given in the transition matrix to calculate transition probabilities we represent the transition table as a matrix $\mathbf{P} = (P_{ij})$, $1 \leq i, j \leq k$, where the entry P_{ij} in the i th row and the j th column represents the transition probability from class i to class j over the time interval, and k is the total number of rating classes (including default). Having added the default class D row ensures that the transition probability matrix is actually a square matrix. Now, to calculate transition probabilities, we need the following two assumptions:

- *Time-invariance.* The probability of a transition from rating class i at time t to rating class j at time $T > t$ does not depend on the calendar dates t and T but only depends on time via the length $T - t$ of the time interval.
- *Markov property.* Besides the length of the time interval, the probability of a transition from rating class i to rating class j only depends on the rating class that we come from (class i) and the rating class that we go to (class j), and on no other external variables.

The assumption of time-invariance is not an uncommon assumption in statistical analysis. Essentially it says that the future will be like the past. Although common, it is very restrictive as it rules out phenomena such as business cycle effects. Empirically, downgrades are very much more likely in recessions than in boom phases. The second assumption, the Markov property, is also restrictive. It essentially rules out the use of any information beyond the current rating class – all we need to know in order to determine future transition probabilities is the current rating class. Besides disallowing other explanatory variables beyond the rating itself, the Markov property also implies that the history of the obligor's rating is irrelevant as long as the current rating is known. This is in contradiction to empirically observed *rating momentum*. Recently downgraded obligors are much more likely to be downgraded again than other obligors of the same rating which have already been in that rating class for a long time. However, the Markov property implies an absence of rating momentum.

Both assumptions thus contradict empirical observation. In practice, the inaccuracies of these assumptions have to be weighed against the significant simplifications that they allow: under time-invariance and the Markov property we can speak of ‘the’ one-year transition probability matrix \mathbf{P} from which we can calculate the two-year and longer-horizon transition probability matrices. For example, the two-period transition probability matrix is reached as follows.

The probability of going from rating class A to rating class BB over two periods equals the sum of the following probabilities:

- $p_{(A \rightarrow AAA)} p_{(AAA \rightarrow B)}$, the probability of going from A to AAA in the first year times the probability of going from AAA to BB in the second year,
- $+ p_{(A \rightarrow AA)} p_{(AA \rightarrow B)}$, the probability of going from A to AA in the first year times the probability of going from AA to BB in the second year,
- $+ \dots$
- $+ p_{(A \rightarrow D)} p_{(D \rightarrow B)}$, the probability of going from A to D in the first year times the probability of going from D to BB in the second year.

Essentially, we need to sum over all possible paths that the rating can take from A at $t = 0$ to BB at $t = 2$.

Mathematically, $p_{ij}^{(2)} = \sum_{n=1}^K p_{in} p_{nj}$ and the two-year transition matrix is therefore given by

$$\mathbf{P}^{(2)} = \mathbf{P} \times \mathbf{P},$$

that is, the matrix product of the one-year transition matrix with itself. Note that we *need* time-invariance so that the transition probabilities in the second year are the same as the ones in the first year, and the Markov property is used when we calculate the transition probabilities for the second year (i.e. we just multiply the transition probabilities, irrespective of whether we multiply two downgrade probabilities, or a downgrade and an upgrade probability). Similarly, the transition matrix over n years is obtained from the one-year transition matrix by multiplication n times with itself:

$$\mathbf{P}^{(n)} = \mathbf{P}^n. \tag{III.B.4.8}$$

In summary, assuming time-invariance and the Markov property means that the one-year transition data given by the rating agencies can be used to calculate the transition probabilities (and default probabilities) for *all* time horizons, by a simple matrix multiplication.

If transition probabilities are sought for shorter time horizons than the one-year horizon given by the original transition matrix \mathbf{P} , one may try to find a short-term (say, monthly) transition matrix \mathbf{A} , that is consistent with \mathbf{P} , or in other words a $1/n$ -period transition matrix such that $\mathbf{A}^n = \mathbf{P}$. In many cases the problem of finding such a short-term transition matrix usually does not possess an exact solution, but for most transition matrices found in practice highly accurate approximate solutions can be found numerically.

III.B.4.4 Credit Scoring and Internal Rating Models

Because the large public credit rating agencies have traditionally concentrated on larger, US-based corporate bond issuers, the set of obligors covered by public credit rating agencies is only a fraction (albeit an important one) of all obligors that make up a bank's credit portfolio. Important missing categories of obligors are small and medium-sized businesses, many larger European, Japanese and Asian obligors, and of course the whole retail portfolio.

In order to assess the credit risk of such obligors, various statistical methods have been developed. Generally, when it comes to statistical models of default prediction, the choice of the *inputs* seems to be more important than the choice of the particular *methodology*, although both are frequently intertwined. Important variables that have been found to drive the default behaviour of obligors include:

- balance-sheet data capturing the indebtedness of the obligor;
- profits and free cash flows capturing the ability to pay;
- the riskiness (volatility) of the business;
- (if available) market data, such as the firm's market capitalisation;
- macro-economic data capturing the effects of the business environment on the obligor.

III.B.4.4.1 Credit Scoring

One of the earliest published credit scoring model goes back to Altman (1968), and was further developed in Altman *et al.* (1977). Credit scoring models usually rely on accounting ratios like the ones listed in Table III.B.4.3.

Table III.B.4.3: Key accounting ratios

Some key accounting ratios, and the averages of these ratios over the firms in the data set used by Altman (1968). 'Bankrupt' firms defaulted within one year.

Accounting Ratio		Bankrupt	Non-bankrupt
Working Capital / Total Assets	X_1	-6.10%	41.40%
Retained Earnings / Total Assets	X_2	-62.10%	35.50%
Earnings before Interest and Taxes/ Total Assets	X_3	-31.80%	15.30%
Market Capitalisation / Debt	X_4	40.10%	247.70%
Sales / Total Assets	X_5	150%	190%

Altman proposes to calculate for each obligor the score function

$$Z^{(i)} = 1.2 \times X_1^{(i)} + 1.4 \times X_2^{(i)} + 3.3 \times X_3^{(i)} + 0.6 \times X_4^{(i)} + 1.0 \times X_5^{(i)}, \quad (\text{III.B.4.9})$$

the value of which is the so-called *Z-score* of the *i*th obligor. The variables $X_1^{(i)}, \dots, X_5^{(i)}$ are the values of the accounting ratios for the *i*th obligor. Table III.B.4.3 shows the definitions of these ratios and their averages in Altman's data set. Altman proposes to use the *Z-score* to make loan decisions. If the score function has a value less than the cutoff score of 1.8 then the obligor is likely to default and a loan is to be denied. The score can also be used to *rank* obligors: a higher score is a sign of better credit quality.

The variables that are used to explain the default behaviour of the obligor are similar for most internal credit scoring models. Indeed, accounting ratios are typically included to measure such things as:

- indebtedness (for instance, using the ratio of market capitalisation to debt),
- cash flow available for debt service (for instance, via earnings before interest and taxes),
- profitability (for instance, using the retained earnings to total asset ratio),

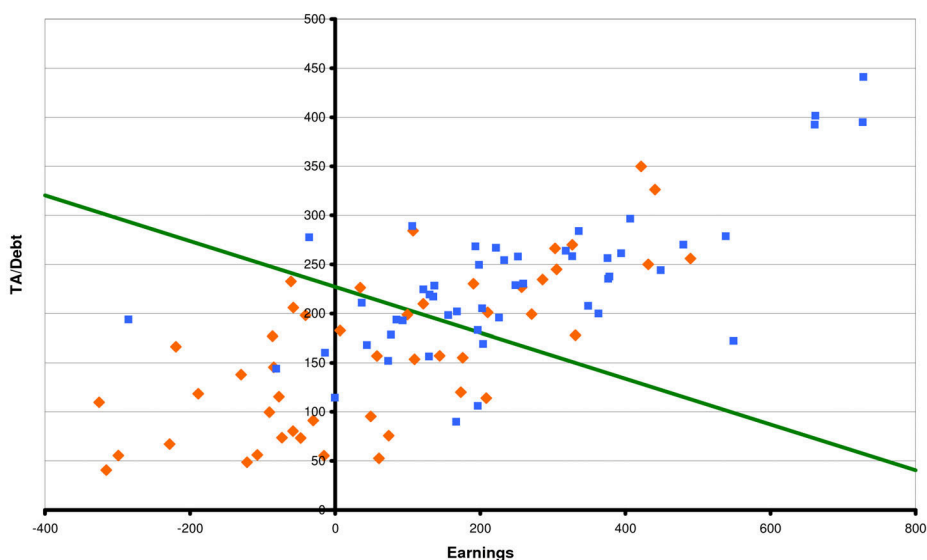
as well as numbers to measure earnings stability and the debt service load.

The scoring weights (1.2, 1.4, 3.3, 0.6, 1.0) and the cutoff level of 1.8 in (III.B.4.9) were estimated using a statistical method. The 'training set' lists the values of each of the variables X_i for a set of obligors and, of course, the information whether the obligors eventually defaulted or whether they survived. When choosing the training set it is important to take care that it is diverse, that it is representative of the real world, and that it contains a sufficient number of *defaulted* obligors. In particular, one has to be very careful if only *internal* data of a bank is used to estimate a scoring

model because this data set will only contain obligors who have already been pre-selected according to the existing lending criteria. If, for example, the existing criteria are very strict regarding earnings before interest and taxes (EBIT), then the internal data will contain very few obligors with bad EBIT numbers and the scoring model may mistakenly conclude that EBIT was not relevant to default prediction.

Figure III.B.4.4: An example of credit scoring

*Defaulted obligors are shown as red points, obligors which survived as blue points.
The green line shows the points where the score is exactly at the cutoff level.*



The scoring weights and the cutoff level in (III.B.4.9) are chosen to maximise the number of correctly classified obligors. Figure III.B.4.4 shows the principle for the case with only two accounting ratios. The red cloud of points are the defaulted obligors in the training set, the blue cloud are the surviving obligors. It is not surprising that we have more and more survivors, the further we go into the ‘northeast’ direction of higher earnings and lower indebtedness. For a given cutoff level, the score weights define a line in this graph which consists of the points that have a score of exactly the cutoff score. In Figure III.B.4.4, the green line shows this line for the optimal cutoff. Any points above that line will have a higher score than the cutoff (and thus will be accepted), any points below it will have a lower score (and will be rejected).

III.B.4.4.2 Estimation of the Probability of Default

The two most common models for the estimation of default probabilities are logit and probit models. In the *probit* model, it is assumed that the default probability of obligor i can be written as:

$$p_i = \Phi(\beta_0 + \beta_1 \times X_1^{(i)} + \dots + \beta_N \times X_N^{(i)}), \quad (\text{III.B.4.10})$$

where $\Phi(\cdot)$ is the cumulative normal distribution function, β_n are parameters that are estimated statistically, and $X_n^{(i)}$ are given accounting ratios and other explanatory variables for obligor i , with $n = 1, \dots, N$.

The structure of equation (III.B.4.10) is very similar to a scoring model like (III.B.4.9). Essentially, we are mapping from the Z-scores to default probabilities by using the cumulative normal distribution function. However, note that the scores have a different interpretation in the two models:

- In the probit model (III.B.4.10) high scores are mapped to high default probabilities.
- In a scoring model such as (III.B.4.9) high scores are mapped to low default risk.

Thus high scores are ‘good’ in the scoring model, but ‘bad’ in the probit model. The estimated parameters of the two models are therefore quite different, even when the same explanatory variables used in the two models.

One advantage of probit models is that we can tell a story about how defaults occur in this set-up. Let us define the *credit index* for the i th obligor as

$$-(\beta_0 + \beta_1 \times X_1^{(i)} + \dots + \beta_N \times X_N^{(i)}) + \varepsilon^{(i)}, \quad (\text{III.B.4.11})$$

where $\varepsilon^{(i)}$ is a standard normally distributed noise component. If we assume that obligor i defaults if his credit index drops below zero, then the default probability of obligor i is exactly equal to the p_i given by the probit model. Thus, there is a natural link from probit models to credit portfolio models like CreditMetrics (see Chapter III.B.5).

The logit model differs from the probit model only in that it does not use the cumulative normal distribution function to map from scores to default probabilities but uses a different function, the *logistic transformation*. Here, the default probability of obligor i is set equal to

$$p_i = \frac{1}{1 + \exp\{\beta_0 + \beta_1 \times X_1^{(i)} + \dots + \beta_N \times X_N^{(i)}\}}. \quad (\text{III.B.4.12})$$

The logit model is slightly easier to estimate than the probit, but generally results from logit and probit models do not differ much.

III.B.4.4.3 Other Methods to Determine the Probability of Default

Besides the rating models presented above, a number of other models are used to assess the credit quality of individual obligors. The most important of these is the KMV model. This is explained in detail in Section III.B.5.6.

In a gross simplification, the KMV model can also be viewed as a scoring model where the ‘accounting ratio’ is the *distance-to-default*. But, this approach is special because the distance-to-default contains both *market* information (the market capitalisation of the firm) and *volatility* information. Both market and volatility information are usually not found in classical accounting ratios. From a purely practical point of view, the large historical database underlying the model is valuable, in particular the empirical transformation that KMV uses to reach a default probability estimate for a given distance-to-default.

Another class of potentially powerful models for credit classification are neural networks and other expert systems from artificial intelligence research. Generally, such models have been found to be equally powerful compared to good statistical models (like logit or probit models), but their acceptance in practice has been hindered by their complex nature, which is essentially a ‘black box’ to the credit officer.

III.B.4.5 Market-Implied Default Probabilities

The credit default tree presented in Figure III.B.4.2 is a perfectly adequate model to price simple credit-sensitive instruments such as corporate bonds or credit default swaps (CDSs), provided that the necessary conditional default probabilities are already specified. In this section we will turn the model around. Instead of determining *prices* for given set of *parameters*, we now determine *parameters* for the *prices* of a set of benchmark securities, the *calibration securities*. The default probabilities that are reached in this calibration exercise are called *market-implied default probabilities*.

This approach rests on several assumptions. First, we assume that there is information to be found in the prices. The prices of the defaultable bonds/CDSs must be meaningful, that is to say, liquid and not unduly affected by external factors beyond default risk (such as taxes). Also, the peculiarities of different markets enable the participants to incorporate their information into the prices to different degrees. For example, while it is easy to short credit risk in the CDS market, it can be difficult to short defaultable bonds. Therefore, calibration securities should be taken from markets with similar conditions.

Second, it is necessary that all calibration instruments are subject to the same type of credit risk, that is, they reference the same obligor (and have the same default definition in the case of CDSs), and they have the same recovery rate and seniority class in default. We also need to know the value of the recovery rate (or at least its average).

Third, we use the risk-neutral pricing paradigm, that is, we price a security by taking the discounted expected value of its payoffs. Thus, the probabilities that we will reach are *pricing* or *martingale measure* probabilities. It is important to realise that these are different from historical probabilities because they are loaded with risk premia. This issue will be discussed in the next subsection.

Finally, we assume that movements of risk-free interest rates and defaults are independent. This is mostly a technical assumption, which significantly simplifies the analysis. In normal situations this correlation has only a second-order effect on the resulting default probabilities.

III.B.4.5.1 Pricing the Calibration Securities

As a running example in this section we consider the problem of calibrating a term structure of default probabilities to the bond prices shown in Table III.B.4.4.

Table III.B.4.4: Calibration securities

Prices of five corporate bonds issued by Daimler Chrysler NA Holding. Trade date: 17 November 2003, effective date 19 November 2003. Coupons are paid annually, currency is USD, notional is 100.

No.	Dirty Price	Coupon	Maturity
1	105.46	4.5	03-01-2005
2	106	5.75	23-06-2005
3	105.27	4.62	10-03-2006
4	100.84	3.75	02-10-2006
5	109.46	5.62	16-01-2007

We can represent the prices of the calibration securities with two elementary types of (hypothetical) building-block securities: defaultable zero-coupon bonds which only pay in survival, and a recovery security which pays at the time of default. Specifically, we denote by $\bar{B}(0, T)$ the value at time $t = 0$ of receiving \$1 at time T in survival (and nothing if default occurs before T), and by $E(0, T)$ the value at time $t = 0$ of receiving \$1 at default, if the default occurs before time T .

The value of a defaultable coupon bond with coupon payment dates T_k , $k = 1, \dots, K$, coupon amount c and recovery rate R can now be written as⁶

$$V^{\text{Bond}} = c\bar{B}(0, T_1) + c\bar{B}(0, T_2) + \dots + (1 + c)\bar{B}(0, T_K) + R \cdot E(0, T_K). \quad (\text{III.B.4.13})$$

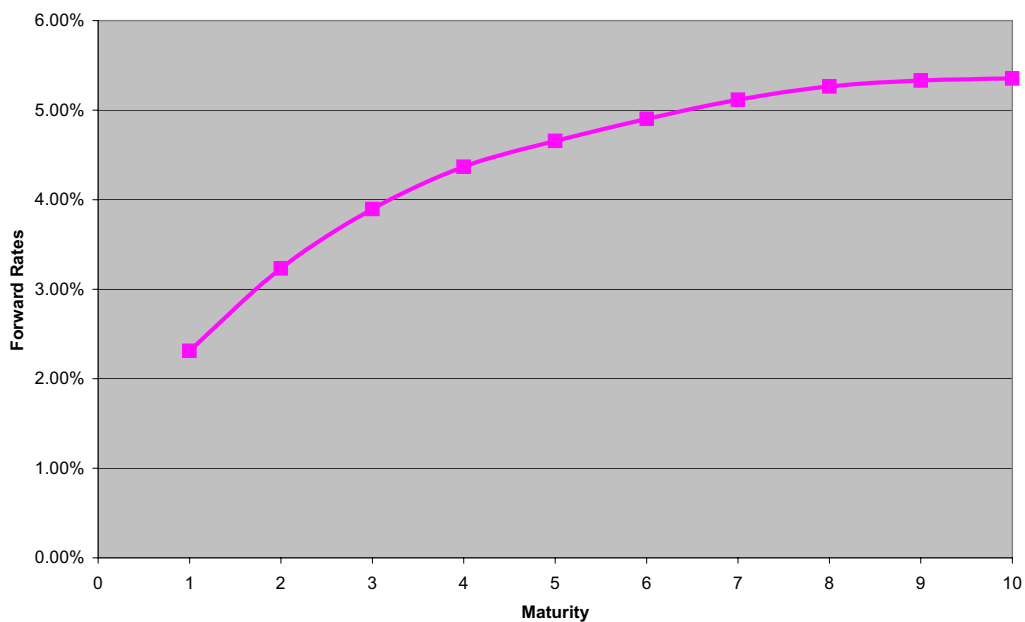
Here, the first summation represents the value of the promised coupon payments and the final repayment of principal, and the final term represents the value of the recovery received at default, with R denoting the recovery rate.

For example, the first bond in Table III.B.4.4 has coupon payments of $c = \$2.25$ at $T_1 = 3$ January 2004 and $T_2 = 3$ January 2005, and the principal repayment of $\$100$ at T_2 . Furthermore, we assume that the recovery rate is 40%; thus we will have an additional payoff of $R = 40$ at default, if a default occurs before T_2 .

Similarly, we can represent the prices of the other four bonds as discounted values of the principal, coupons and recovery cash flows. For all bonds together, we have promised cash flows on the following payment dates: 3 January 2005, 16 January 2005, 10 March 2005, 23 June 2005, 2 October 2005, 19 November 2005, and 3 January 2006. These are the maturities of the defaultable zero-coupon bonds that we need to represent the cash flows of our calibration securities.

Figure III.B.4.5: The term structure of risk-free interest rates

USD forward Libor rates on 17 November 2003.



⁶We ignore day count conventions and other technical adjustments. Notional amounts are normalised to 1.

The value of a protection-buyer position in a CDS on the reference credit with CDS rate \bar{s} and the same payment dates $T_k, k = 1, \dots, K$, is:

$$V^{\text{CDS}} = -\bar{s} \left(\bar{B}(0, T_1) + \dots + \bar{B}(0, T_N) \right) + (1 - R) \times E(0, T_K). \quad (\text{III.B.4.14})$$

Here, the first sum represents the value of the fee stream (a liability to the protection buyer), and the last term represents the value of receiving $1 - R$ at default of the obligor. The market CDS rate is chosen such that the value of the CDS position is zero:

$$\bar{s} = (1 - R) \frac{E(0, T_K)}{\bar{B}(0, T_1) + \dots + \bar{B}(0, T_N)}. \quad (\text{III.B.4.15})$$

Having reduced the pricing problems to the problem of finding prices for $\bar{B}(0, T)$ and $E(0, T)$, we now have to represent these prices in terms of the survival probabilities defined in Section III.B.4.1. The price of a defaultable zero coupon bond is easily seen to be

$$\bar{B}(0, T) = B(0, T) \cdot P(0, T), \quad (\text{III.B.4.16})$$

where $B(0, T)$ is the price of a default-free zero-coupon bond with maturity T , and $P(0, T)$ the survival probability until T . It can also be shown that:

$$E(0, T) = p_1 B(0, T_1) + p_2 P(0, T_1) B(0, T_2) + \dots + p_k P(0, T_{k-1}) B(0, T_k). \quad (\text{III.B.4.17})$$

Here, $p_k P(0, T_{k-1})$ represents the probability of surviving until time T_{k-1} (which is $P(0, T_{k-1})$), and then defaulting in the period $[T_{k-1}, T_k]$ (which occurs with probability p_k). This takes us to the default branch at time T_k . Then, after discounting with $B(0, T_k)$, we reach the formula above.

Example III.B.4.2

In the example given above, we use the USD forward Libor curve as the risk-free interest rates. This yields the prices of the default-free zero-coupon bonds $B(0, T)$ at the payment dates of the calibration bonds. In equations (III.B.4.7) and (III.B.4.8) we also need default and survival probabilities for several different time horizons. Here we made the assumption that these probabilities are given by a constant default-intensity model as in Section III.B.4.1, that is, that $P(T) = \exp\{-\lambda_0 T\}$, where λ_0 is a parameter that we need to find. Finally, we assume a common recovery rate of 40%.

Given all these assumptions, the only remaining degree of freedom that we can use to match model prices to market prices is the default intensity λ_0 . By using Excel's Solver routine we find that a value of $\lambda_0 = 1.0029\%$ minimises the squared pricing errors of the bonds' model prices

relative to the market prices. The results are presented in Table III.B.4.5. The default payoffs row shows the values of the potential recovery payoffs. The survival payoffs row shows the values of the promised coupon and principal payments of the bonds. The sum of these two yields the model price of the bond.

Table III.B.4.5: Model prices of the calibration securities

The model prices of the bonds of Table III.B.4.4 under the assumption of a constant default intensity $\lambda_0 = 1.0029\%$, a recovery rate of 40% and using the default-free interest rates shown in Figure III.B.4.5.

Coupon	4.5	5.75	4.62	3.75	5.62
Maturity	03-01-2005	23-06-2005	10-03-2006	02-10-2006	16-01-2007
Market Price	105.46	106	105.27	100.84	109.46
Calibration Model Prices					
Default Payoff	0.4482	0.6289	0.8966	1.1020	1.2055
Survival Payoffs	105.0228	105.3767	104.5247	99.2955	108.5792
Model Price	105.4710	106.0056	105.4212	100.3975	109.7847

We have introduced a variety of ways to represent default probabilities in Section III.B.4.1. Similarly, there are several ways to represent the prices of defaultable securities. A particularly simple representation can be reached for the fair CDS rate (III.B.4.15) if the ‘odds’ of default H_k are used:

$$\bar{s} = (1 - R) \times (w_1 H_1 + \dots + w_K H_K). \quad (\text{III.B.4.18})$$

That is, the CDS rate \bar{s} equals the loss on default $(1 - R)$ times a weighted average of the odds of default H_k where the weights of the average are

$$w_k = \frac{\bar{B}(0, T_k)}{\bar{B}(0, T_1) + \dots + \bar{B}(0, T_K)}.$$

Clearly, these weights are non-negative and sum to 1.

In the particularly simple case of constant odds of default (i.e. if all H_k take the same value H) we have

$$\bar{s} = (1 - R)H. \quad (\text{III.B.4.19})$$

Hence, if we equate the odds of default with the default hazard rate (a very accurate approximation) we can say that *the CDS rate equals the loss given default times the default hazard rate.*

III.B.4.5.2 Calculating implied default probabilities

Backing out an *implied* default hazard rate from a single CDS quote is very straightforward, given equation (III.B.4.19). If we observe a CDS spread \bar{s} in the market, then the corresponding implied default hazard rate is reached by solving equation (III.B.4.19) for the hazard rate:

$$\hat{H} = \bar{s} / (1 - R).$$

Example III.B.4.3

In the case of Daimler-Chrysler, the quote for a CDS with five years' maturity on 17 November 2003 was 108.13bp. According to the formula given above, the implied default hazard rate is 1.8% (at an assumed recovery rate of 40%). It is not unusual that implied default intensities from CDSs are higher than the corresponding implied default intensities from bond prices. This effect is partly due to the fact that the embedded delivery option of a CDS makes the recovery rate with a CDS smaller than the recovery rate of a bond, and partly caused by market imperfections in the bond markets.

In a general situation, we may have CDS quotes for several maturities or we may have market prices for different bonds with different maturities and different coupons. In order to find a set of implied default probabilities that is consistent with these prices and that simultaneously looks 'sensible', a numerical optimisation routine must be used. This procedure is quite similar to the bootstrapping procedures used to back out term structures of interest rates from default-free bond prices. More details can be found in Schönbucher (2003). Generally, apart from introducing a time-dependence in the default probabilities, the results are qualitatively similar to the simple result for the single CDS. That is, credit spreads are approximately equal to the default hazard rate times the loss given default.

III.B.4.6 Credit Rating and Credit Spreads

In order to compare implied default probabilities and historical default probabilities, it is instructive to study Figure III.B.4.6. In particular, let us take the year 1997. For that year, the dark blue line shows an average credit spread for US Baa-rated corporate bonds of 0.65%, which will yield an implied default hazard rate of 1.08% (assuming a recovery rate of 40%). This spread is measured as spread over Aaa-rated corporate bonds and not as spread over treasuries (in order to avoid tax and liquidity premia, which would be present in the treasury market).⁷

⁷Choosing Aaa instead of risk-free only makes our estimate more conservative.

The spread is a compensation for the default risk that we bear if we invest in Baa-rated bonds, so it is interesting to see if this compensation was adequate. If the realised default hazard rate equals the implied default hazard rate, then an investment in Baa-rated bonds will just break even. This is what we would expect to happen with risk-neutral investors and in the absence of market imperfections. We now investigate how the implied default intensities compare, historically, to actual default rates, and whether the compensation through credit spread was adequate for the credit risk incurred.

In fact, it turns out that the compensation is much more than adequate. Imagine we had bought the total Baa-rated corporate bonds market (or an equally weighted fraction of it) in the year 1997 and held it to maturity – let us call this portfolio ‘Baa97’. The value of the yellow line in 1997 tells us the rate of defaults that we would have suffered over one year, that is, 1997–1998. Clearly, this rate of defaults is almost zero!

So, what if we had to hold the Baa97 portfolio for three years starting in 1997?⁸ The blue line tells us that over 1997–2000 the annual default rate of the Baa97 bond portfolio was only 0.3%. This is still much less than the implied default rate, and even less than the spread.

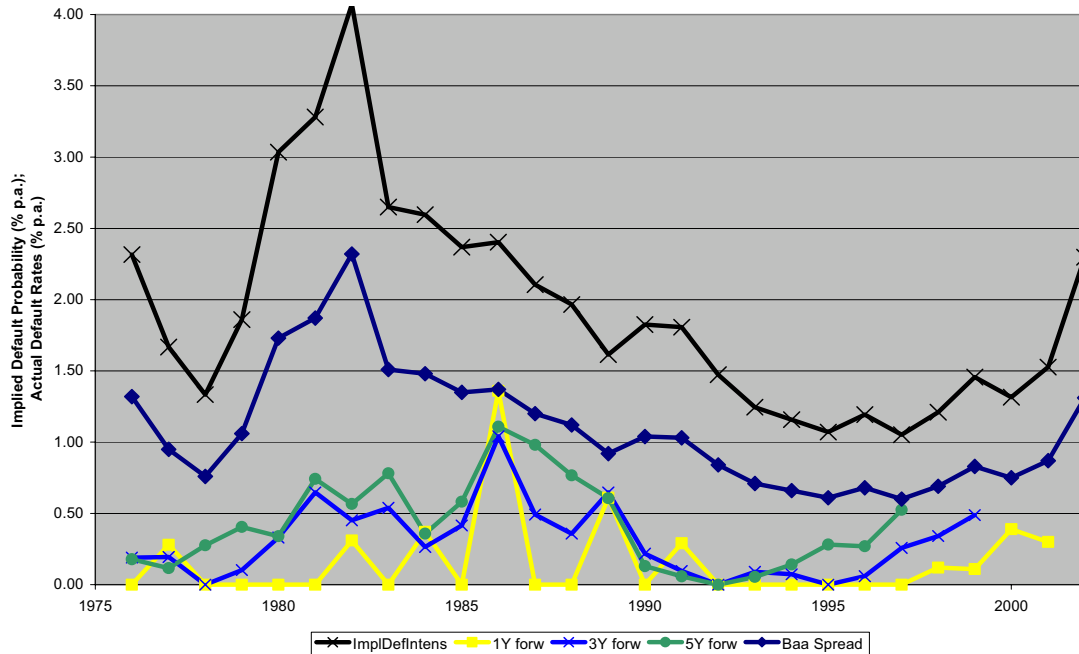
Even over a five-year horizon starting 1997 the historical default rate is only about 0.55% p.a. (see the value of the yellow line in 1997), much below the implied default rate. This situation repeats consistently without exception over the whole period from 1976 to 1997 (1997 is the last year in which we could calculate a five-year forward-looking default rate).

Hence, investing in Baa portfolios would *always and without exception* have outperformed an investment in the risk-free reference portfolio (which here is Aaa-rated bonds).

⁸We want to hold until maturity in order to eliminate effects due to market price movements.

Figure III.B.4.6: Implied default probabilities vs. historical default frequencies

This figure shows, for the pool of Baa-rated US corporate bonds: implied default rates (black, for 40% recovery), credit spreads (dark blue, measured as spreads over Aaa, not as spreads over treasury), the default rate of the pool over the next year (yellow), over the next three years (blue), and over the next five years (green).



Other studies also indicate that there seems to be a significant discrepancy between implied default rates and historical default rates. Typically, implied default rates tend to be larger by a factor of 2–3. This situation has been termed the *spread premium puzzle*: why are spreads so much higher than seems to be justified by their actual credit risk component? Or is this an arbitrage opportunity?

Several explanations have been put forward.

- First, there is risk aversion. An investment in Baa97 will lose money if a recession comes, but a recession is exactly the situation in which the average investor needs money. Therefore, the investor will demand a higher return in order to be compensated for this wrong-sided risk.
- Secondly, maybe the actual default risk was much higher than historical incidence, we just did not experience the truly bad scenario, but the possibility of this bad scenario was nevertheless priced into the spreads.
- Third, maybe there were tax effects at work that made the Baa-rated bonds unattractive, or liquidity premia were demanded for investments in Baa97.

Unfortunately, none of these explanations can fully explain the effect. Risk aversion is certainly present, but risk premia can never lead to a shift in spreads that amounts to a virtual arbitrage opportunity: that would be inconsistent with rational investor behaviour because even the most risk-averse investor could have picked up a seemingly risk-free excess return here.

The second explanation concerning the unrealised ‘Armageddon’ scenario also cannot explain the size of the effect. If this mysterious extremely bad scenario were to explain a sizeable proportion of the spreads, then it must also have a probability that is not too small compared to the individual default probability of an obligor. But if this is the case, why has this scenario never occurred so far?

The liquidity argument is certainly valid if spreads over treasuries are considered. But here we are considering spreads over Aaa corporates which should have similar liquidity problems to Baa-rated bonds.

Finally, the tax argument again does not hold for spreads over Aaa, it only has the possibility of being relevant if spreads over treasuries are considered, because treasuries are exempt from state taxes in the USA. Besides, the *spread premium puzzle* is also observed in markets which are not affected by US tax rules, for example in the markets for Eurobonds or for bonds of non-US issuers.

So the spread-premium puzzle remains a puzzle. Maybe it is indeed a market imperfection. But maybe it has already disappeared: since mid-2003 spreads have decreased significantly compared to the spreads used in Figure III.B.4.6. It is quite possible that the market has finally reached a level where implied and actual default risk are approximately equal, or at least in a more realistic relationship to each other. Of course, we will only know this for sure after it is too late to make an investment decision.

III.B.4.7 Summary

Credit rating and the estimation and measurement of default probabilities are a classical problem of credit analysis, and one that never seems to be perfectly solvable. The classical solution to this problem is to rely on the rating assessment of an external rating agency – essentially, this is reliance on expert advice. We have seen how these rating classifications can be translated into concrete numbers for default and survival probabilities over different time horizons.

Recent advances in computing power and (more importantly) the increasing availability of the necessary data in electronic form have made it possible to estimate default probabilities on a

purely statistical basis. In many cases, such quantitative approaches are able to compete successfully with agency ratings, and frequently they are the only option when no agency rating is available.

Finally, we discussed methods to imply default probabilities from observed market prices of traded credit-sensitive instruments such as bonds and credit default swaps. While these methods are indispensable to assess the risk compensation that one should get for the type of credit risk under consideration, spread-implied probabilities usually differ significantly and systematically from statistical and historical default rates. This credit-spread puzzle remains an open question to date. Nevertheless, because they are systematically above historical default rates, spread-implied probabilities may be useful as very conservative estimates of default probabilities.

In practice, implied probabilities are the correct probabilities to use for pricing applications, because these probabilities already contain the risk premia that are paid for the credit risk contained in the calibration securities. Risk management, capital allocation and value-at-risk calculations, on the other hand, require historical probabilities because here the preferences and risk aversion can be added later on.

References

Altman, E (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, **23**(4), pp. 589–609.

Altman, E, Haldeman, R, and Narayanan, P (1977) Zeta analysis: a new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance*, **1**, pp. 31–54.

Hamilton, D T, Cantor, R, and Ou, S (2002) Default and recovery rates of corporate bond issuers. Special comment, Moody's Investor Service Global Credit Research, February.

Schönbucher, P J (2003) *Credit Derivatives Pricing Models*. Chichester: Wiley.

III.B.5 Portfolio Models of Credit Loss

Michel Crouhy, Dan Galai and Robert Mark¹

This chapter describes the main approaches to the modelling of credit risk in a portfolio context (credit value-at-risk), i.e. the credit migration approach, the contingent claim or structural approach, and the actuarial approach. It reviews the assumptions of the credit portfolio models and the pros and cons of each approach. Finally, it discusses the relationship between credit value-at-risk, economic capital and regulatory capital.

III.B.5.1 Introduction

In this chapter we review the main approaches to modelling credit risk. For each approach we explain the basic logic behind it, describe the data required and evaluate its strengths and weaknesses. The interested reader can find a more detailed description of the approaches in Crouhy *et al.* (2001).

A bank should be concerned with the estimation of the risk of default of a specific creditor, since this is the basis for pricing a loan and charging the borrower with the appropriate interest rate. But, at the same time, the bank should be looking at the quality of its loan portfolio as a whole, since the stability of the bank depends to a large extent on the performance of its portfolio, and on the size of credit-related losses in the portfolio in a given period. Portfolio analysis may in turn affect the pricing of individual loans and the lending decision as each asset's contribution to portfolio risk must be considered.

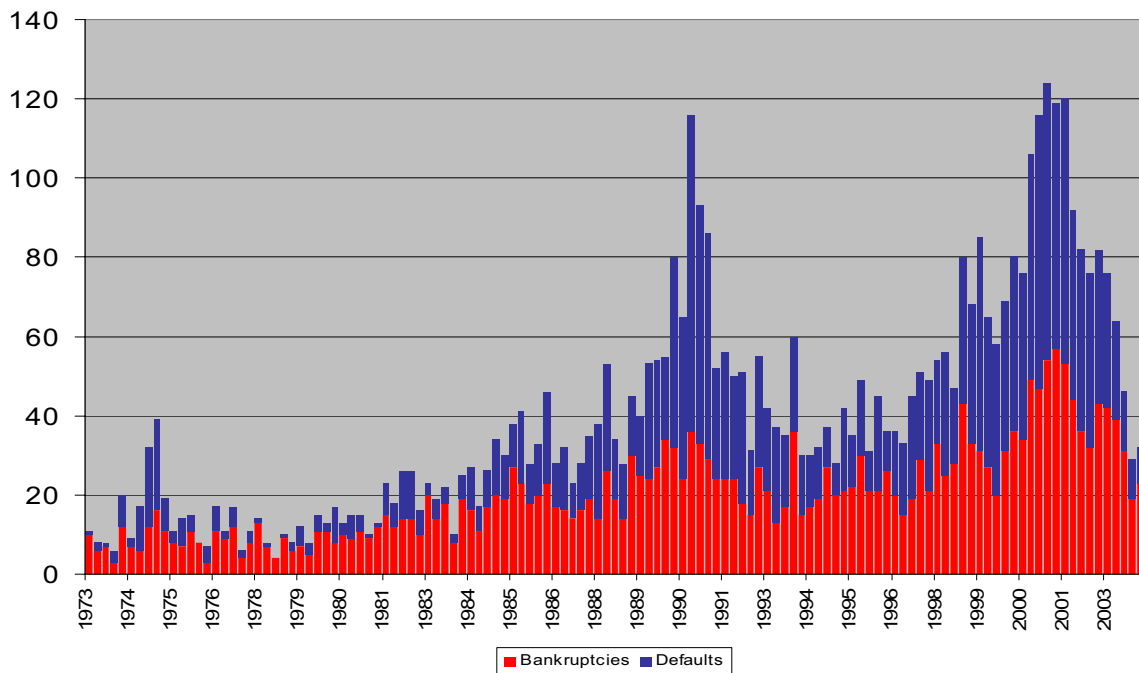
Modelling credit risk and pricing risky loans or bonds is a complicated task. The factors that affect credit risk are many and varied. Some factors are exogenous or economy-wide, such as the level of interest rates and the growth rate of the economy. Other factors are endogenous, such as the business risk of the firm, its capital structure, and the flexibility of its production technology. A major consideration is whether to evaluate credit risk as a discrete event, and concentrate only on the potential default event, or whether to analyse the dynamics of the debt value and the associated credit spread, and to estimate its risk over the whole time interval to its maturity. Another important issue is the data sources that are available in order to assess credit risk. Can the analyst rely on accounting data, or are these too stale and subject to manipulation? To what extent are market data available, and then, to what extent are the markets efficient enough to convey reliable information?

¹ Michel Crouhy is a Partner at Black Diamond, Dan Galai is a Professor at the Hebrew University and Principal at Sigma P.C.M., and Robert Mark is CEO of Black Diamond.

Before proceeding further, let us define some fundamental concepts. Default, in theory, occurs when the asset value falls below the value of the firm's liabilities (Merton, 1974). Default, however, is distinct from bankruptcy. Bankruptcy describes the situation in which the firm is liquidated, and the proceeds from the asset sale are distributed to the various claim holders according to pre-specified priority rules. Default, on the other hand, is usually defined as the event that a firm misses a payment on a coupon and/or the reimbursement of principal at debt maturity. Cross-default clauses on debt contracts are such that when the firm misses a single payment on a debt, it is declared in default on all its obligations.

Since the early 1980s, Chapter 11 regulation in the United States has protected firms in default and helped to maintain them as going concerns during a period in which they attempt to restructure their activities and their financial structure. Figure III.B.5.1 compares the number of bankruptcies to the number of defaults during the period from 1973 to 2004. The data are for North American public companies, but note that legal procedures in enforcing the bankruptcy procedures in the case of a default event vary quite substantially across jurisdictions (see J.P. Morgan, 1997, Appendix G).

**Figure III.B.5.1: Bankruptcies and defaults in North American public companies
1973Q1 to 2004Q1**



Over the last few years, a number of new approaches to credit risk modelling have been made public. The CreditMetrics approach (which was initiated by J.P.Morgan and was spun off to RiskMetrics Inc.) is based on the analysis of credit migration, i.e. the probability of moving from one credit grade to another, including default, within a given time horizon which is usually one year. CreditMetrics estimates the full, one-year forward distribution of the values of any bond or loan portfolio, where the changes in values are related to credit migration only. The past migration history of thousands of rated bonds is assumed to accurately describe the probability of migration in the next period. The credit migration framework is reviewed in Section III.B.5.3.

Tom Wilson (1997a, 1997b) proposes an improvement to the credit migration approach, CreditPortfolioView, by allowing default probabilities to vary with the credit cycle. In this approach, default probabilities are a function of macro-variables such as unemployment, the level of interest rates, the growth rate in the economy, government expenses and foreign exchange rates. These macro-variables are the factors which, to a large extent, drive credit cycles. This methodology is reviewed in Section III.B.5.4.

The structural approach to modelling portfolio credit risk offers an alternative to the credit migration approach. Here, the economic value of default is presented as a put option on the value of the firm's assets. The contingent claim approach is introduced in Section III.B.5.5

KMV Corporation, a firm that specialises in credit risk analysis, has developed a credit risk methodology and extensive database to assess default probabilities and the loss distribution related to both default and migration risks. KMV's methodology differs from CreditMetrics in that it relies upon the 'expected default frequency' for each issuer, rather than upon the average historical transition frequencies produced by the rating agencies for each credit class. The KMV approach is based on the asset value model originally proposed by Merton (1974). KMV's methodology, together with the contingent claim approach to measuring credit risk, is reviewed in Section III.B.5.6.

At the end of 1997, Credit Suisse Financial Products released CreditRisk+, an approach that is based on actuarial science. CreditRisk+, which focuses on default alone rather than credit migration, is examined briefly in Section III.B.5.7. CreditRisk+ makes assumptions concerning the dynamics of default for individual bonds or loans, but ignores the causes of default, contrary to KMV and CreditMetrics.

III.B.5.2 What Actually Drives Credit Risk at the Portfolio Level?

Banks, as regulated institutions, are very focused on the quality of their credit portfolio. Banks must assign regulatory capital against credit risk. The current regulation requires banks to assign regulatory capital against each loan obligation, usually 8% of the principal amount. Future regulation, as described in Chapter III.B.6, will allow for better differentiation among obligors based on their ratings. However, the regulators will also look at the quality of the loan portfolio, and the level of concentration by industry and region ('Pillar II' in the New Basel Accord).

But beyond the formal regulatory requirements, banks are judged and evaluated by their shareholders, as well as by their customers, especially the depositors. Therefore, banks have strong incentives to monitor the risk of their assets, and in particular the risk of their loan portfolio. The profitability of most banks largely depends on the performance of the loans they granted in the past. So what are the major factors that affect the performance of the loan portfolio? It should be emphasised that performance has (at least) two dimensions: return and risk. The risk of a loan portfolio can be tricky to assess since a bank can show a nice profitability over a few years due to high interest charges and low default rates, and then, once a default event (or events) occurs on a major exposure, the bank can incur a substantial loss, instantaneously wiping out those profits.

The first factor affecting the portfolio is the credit standing of individual obligors. One bank may concentrate on prime, investment-grade obligors, granting loans only to the best credits, with very low probability of default for any obligor. Another may choose to concentrate on riskier, speculative-grade obligors who pay a much higher coupon rate on their debt. The critical issue for both types of institution is to charge the appropriate interest rate to each borrower that compensates the lender for the risk it undertakes.

The second factor is 'concentration risk', or the extent to which the obligors are diversified across geography and industries. A bank with corporate clients mostly in commercial real estate is considered to be riskier than a bank with corporate loans distributed over many industries. Also, a bank serving only a narrow geographical area can be devastated by a slowdown in the economic activity of that particular region.

This leads to a third important factor that affects the risk of the portfolio: the state of the economy. During good times of economic growth the frequency of defaults falls sharply compared to periods of recession. There is a propensity for things to go wrong at the same time, usually at the trough of the economic cycle. In addition, periods of high default rates such as 2001–2002 are characterised by low recoveries that lead to high loss rates.

The quality of the portfolio can also be affected by the maturity of the loans. Usually, longer loans are considered riskier than short-term loans. Time diversification can reduce the risk of the portfolio by spreading maturities over the economic cycle, as well as reducing ‘liquidity risk’. Liquidity risk is defined as the risk that the bank will run into difficulties when refinancing its assets, for instance by renewing deposits or by raising money through issuing debt instruments, because the market ‘dried up’ or prices increased sharply.

Risk assessment of the portfolio is needed to determine how much economic capital should be allocated against unexpected credit losses. Therefore, the future distribution of the values of the loan portfolio must be estimated. This task is not at all straightforward and is much more complicated than estimating the value of a portfolio of market traded instruments such as stocks and bonds. The major obstacle lies in the estimation of the correlations among potential default events. While we have a lot of data on market traded instruments, we do not have data on non-traded debt instruments. The data problem is also aggravated by statistical issues, for instance that default correlations are not directly observable.

To overcome some of the estimation problems, most approaches imply default correlations from equity correlations as in Section III.B.5.3.2. Still, the estimation problem is huge since so many pairs of cross-correlations must be estimated for a portfolio of obligors. For example, a small portfolio of 1000 obligors requires the estimation of $1000 \times 999 / 2 = 499,500$ correlations. This last problem is circumvented by using a multi-factor or a multi-index statistical model. The rate of return for each firm or stock is assumed to be generated by a linear combination of a few indices. For example, the indices can be related to a country or an industry. This approach reduces the calculation requirements to merely estimating the correlations among pairs of indices.

All the simplifying assumptions are used in order to estimate the portfolio’s credit ‘value-at-risk’ (credit VaR). The distribution of the rate of return of the portfolio of obligors is estimated, and the credit VaR is derived from a percentile of that distribution. The credit VaR of a loan portfolio is thus derived in a similar fashion to market risk, except that the risk horizon is usually much longer. It is simply the distance from the mean to the percentile of the forward distribution, at the desired confidence level. However, the future point in time is typically one year for both regulatory and economic credit risk capital, whereas for market VaR the risk horizon is 10 days for regulatory capital (but again, usually one year for economic capital).²

² The choice of a risk horizon is somewhat arbitrary. It is usually one year as it corresponds to the planning cycle and the average time it would require to recapitalise the bank if it were to suffer a major unexpected loss.

Economic capital is the financial cushion that a bank uses to absorb unexpected losses, including those related to credit events such as credit migration and/or default (see Chapter III.B.6). Figure III.B.5.2 illustrates how the capital charge related to credit risk can be derived from the portfolio value distribution, using the following notation:

$P(p)$ = value of the portfolio in the worst case scenario at the $p\%$ confidence level

FV = forward value of the portfolio = $V_0 (1 + PR)$

V_0 = current mark-to-market value of the portfolio

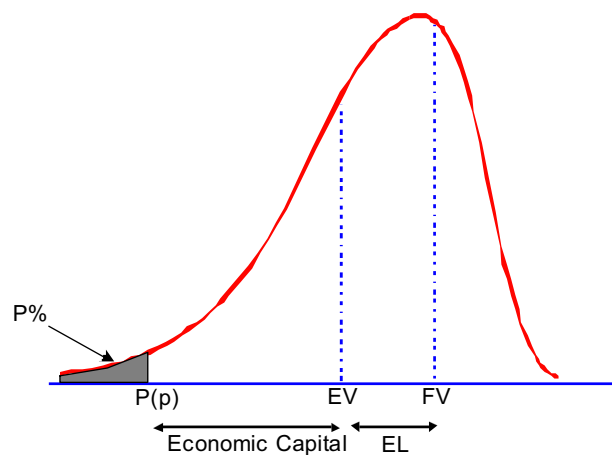
PR = promised return on the portfolio

EV = expected value of the portfolio = $V_0 (1 + ER)$

ER = expected return on the portfolio

EL = expected loss = $FV - EV$.

Figure III.B.5.2: Credit VaR and economic capital attribution



Because the expected loss is priced in the interest charged on loans, it is not part of required economic capital. The capital charge is instead a function of the unexpected losses:

$$\text{Economic Capital} = EV - P(p)$$

When the risk horizon is one year, credit VaR and economic capital are equivalent.

The bank should hold reserves against these unexpected losses at a given confidence level, say 0.01%, so that there is only a 1 in 10,000 chance that the bank will incur losses above the capital level over the period corresponding to the credit risk horizon, say, one year. The choice of a confidence level is generally associated with some target credit rating from a rating agency such as Moody's or Standard and Poor's. Most banks today are targeting a AA debt rating, which implies a probability of default of 3–5 basis points, which then corresponds to a confidence level in the

range of 99.95–99.97%. This confidence level is also the expression of the ‘risk appetite’ of the bank.

III.B.5.3 Credit Migration Framework

Credit migration is a methodology based on the estimation of the forward distribution of changes in the value of a portfolio of loan and bond-type products at a given time horizon, usually one year.³ The changes in value are related to the migration, upwards and downwards, of the credit quality of the obligor, as well as to default. This approach is based on historical data of ratings of many bonds by a rating agency, or on the internal database of a bank. Forward values and exposures are derived from deterministic forward curves of interest rates. The only uncertainty in CreditMetrics relates to credit migration, i.e. the process of moving up or down the credit spectrum. Market risk is ignored in this framework as forward values and exposures are derived from deterministic forward curves.

A typical portfolio distribution is shown Figure III.B.5.3 – it is far from being normal, contrary to market VaR. While it may be reasonable to assume that changes in portfolio values are normally distributed when due to market risk, credit returns are by their nature highly skewed and fat-tailed. An improvement in credit quality brings limited ‘upside’ to an investor, while downgrades or defaults bring with them substantial ‘downsides’. Unlike market VaR, the percentile levels of the distribution cannot be estimated from the mean and variance only. The calculation of VaR for credit risk thus demands a simulation of the full distribution of the changes in the value of the portfolio.

The CreditMetrics risk measurement framework consists of two main building blocks:

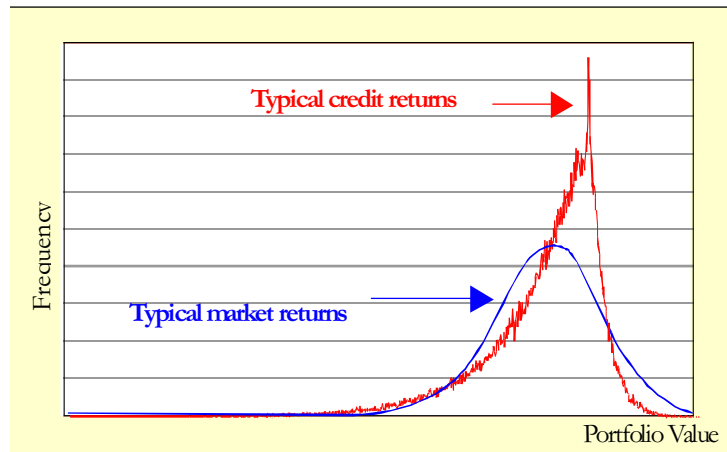
- VaR due to credit for a single financial instrument; and
- VaR at the portfolio level, which accounts for portfolio diversification effects.

The first step is to specify a rating system, with rating categories, together with the probabilities of migrating from one credit quality to another over the credit risk horizon. This *transition matrix* is the key component of the credit migration approach. The matrix may take the form of the historical migration frequencies published by an external rating agency such as Moody’s or Standard & Poor’s, or it may be based on the proprietary rating system internal to the bank. A strong assumption made by CreditMetrics is that all issuers within the same rating class are

³ CreditMetrics’ approach applies primarily to bonds and loans, which are both treated in the same manner. It can be easily extended to any type of financial claims such as receivables, financial letters of credit for which we can easily derive the forward value at the risk horizon for all credit ratings. For derivatives such as swaps or forwards the model needs to be somewhat adjusted or ‘twisted’, since there is no satisfactory way to derive the exposure, and the loss distribution, within the proposed framework (since it assumes deterministic interest rates).

homogeneous credit risks: they have the same transition probabilities and the same default probability.

Figure III.B.5.3: Comparison of the probability distributions of credit returns and market returns



Second, the risk horizon should be specified. This is usually taken to be one year. The third step consists of specifying the forward discount curve at the risk horizon for each credit category. In the case of default, the value of the instrument should be estimated in terms of the ‘recovery rate’, which is given as a percentage of face value or ‘par’. In the final step, this information is translated into the forward distribution of the changes in the portfolio value following credit migration.

III.B.5.3.1 Credit VaR for a Single Bond/Loan

For a given bond in the portfolio, we estimate the distribution of changes in the bond value over a one-year period. Therefore, we have to estimate the assumed values of the bond a year from now for all possible migration events. The most probable event is that the bond will maintain its rating by the end of the year (e.g. a BBB bond has a probability of 86.93% of retaining its BBB rating after a year – see table III.B.5.1) Another migration event is that the bond will be downgraded by one notch, e.g. a BBB bond has a probability of 5.3% of being downgraded to BB within the year. For each credit migration event we use the relevant forward zero-coupon curve, estimated for the ‘new’ possible rating of the bond. These rates serve as discount factors, by which the value of the future cash-flows from the bond are discounted in order to find the value of the bond at the end of the year for the ‘new’ possible rating.

In our example the rating categories, as well as the transition matrix, are chosen from an external rating system (such as the S&P transition matrix in Table III.B.5.1) or an internal rating system.

Table III.B.5.1: Transition matrix: probabilities of credit rating migrating from one rating quality to another, within one year

Initial Rating	Rating at year-end (%)							
	AAA	AA	A	BBB	BB	B	CCC	Default
AAA	90.81	8.33	0.68	0.06	0.12	0	0	0
AA	0.70	90.65	7.79	0.64	0.06	0.14	0.02	0
A	0.09	2.27	91.05	5.52	0.74	0.26	0.01	0.06
BBB	0.02	0.33	5.95	86.93	5.30	1.17	1.12	0.18
BB	0.03	0.14	0.67	7.73	80.53	8.84	1.00	1.06
B	0	0.11	0.24	0.43	6.48	83.46	4.07	5.20
CCC	0.22	0	0.22	1.30	2.38	11.24	64.86	19.79

Source: Standard & Poor's CreditWeek (15 April 1996)

In the case of Standard & Poor's, there are seven rating categories. (It should be noted that the rating agencies supply more granular statistics where each rating category is split into three sub-categories, e.g. A+, A and A– for Standard & Poor's rating category A) The highest category is AAA, the lowest CCC. Default is defined as a situation in which the obligor cannot make a payment related to a bond or a loan obligation, whether the payment is a coupon payment or the redemption of the principal.

The bond issuer in our example currently has a BBB rating. The shaded row in Table III.B.5.1 shows the probability, as estimated by Standard & Poor's, that this BBB issuer will migrate over a period of one year to any one of the eight possible states, including default. Obviously, the most probable situation is that the obligor will remain in the same rating category, BBB; this has a probability of 86.93%. The probability of the issuer defaulting within one year is only 0.18%, while the probability of it being upgraded to AAA is also very small, 0.02%. Such a transition matrix is produced by the rating agencies for all initial ratings, based on the history of credit events that have occurred to the firms rated by those agencies. Moody's publishes similar information.

Although ten or twenty years ago these two rating agencies concentrated on US companies, they now cover tens of thousands of companies around the world. Transition matrices for Japan, Europe and other regions are now becoming available. However there are still some regions where historical default data are insufficient to estimate transition matrices. In such regions the KMV methodology, which does not rely on transition matrices, is often the method of choice. In the US, the transition probabilities published by the agencies are based on more than 20 years of data across all industries. But even these data should be interpreted with care since they only

represent average statistics across a heterogeneous sample of firms, and over several business cycles. For this reason many banks prefer to rely on their own statistics, which relate more closely to the composition of their loan and bond portfolios.

The realised transition and default probabilities also vary quite substantially over the years, depending upon whether the economy is in recession or is expanding (see Section III.B.5.4). When implementing a model that relies on transition probabilities, one may have to adjust the average historical values shown in Table III.B.5.1, to be consistent with one’s assessment of the current economic environment.

A study provided by Moody’s (Carty and Lieberman, 1996) provides some idea of the variability of default rates over time. Historical default statistics (mean and standard deviation) by rating category for the population of obligors that they rated during the period 1970–1995 are shown in Table III.B.5.2. Clearly the default rates become more volatile as credit quality deteriorates. Thus one should expect the elements of the transition matrix corresponding to low grade issuers to change considerably over time, whilst transition probabilities for high grade issuers are unlikely to change much.

Table III.B.5.2: One-year default rates by rating, 1970–1995

Credit rating	One year default rate	
	Average (%)	Standard deviation (%)
Aaa	0.00	0.0
Aa	0.03	0.1
A	0.01	0.0
Baa	0.13	0.3
Ba	1.42	1.3
B	7.62	5.1

Source: Carty and Lieberman (1996)

Now consider the valuation of a bond. This is derived from the zero curve corresponding to the rating of the issuer. Since there are seven possible credit qualities, seven ‘spread’ curves are required to price the bond in all possible states (Table III.B.5.3). All obligors within the same rating class are then marked to market using the same curve. The spot zero curve is used to determine the current spot value of the bond. The forward price of the bond one year from the present is derived from the forward zero curve, one year ahead, which is then applied to the residual cash-flows from year 1 to the maturity of the bond.

Table III.B.5.3: One-year forward zero curves for each credit rating (%)

Category	Year 1	Year 2	Year 3	Year 4
AAA	3.60	4.17	4.73	5.12
AA	3.65	4.22	4.78	5.17
A	3.72	4.32	4.93	5.32
BBB	4.10	4.67	5.25	5.63
BB	5.55	6.02	6.78	7.27
B	6.05	7.02	8.03	8.52
CCC	15.05	15.02	14.03	13.52

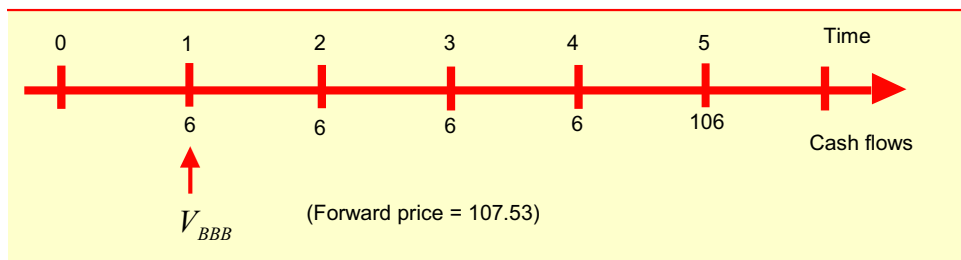
Source: CreditMetrics, J.P. Morgan

From Chapter I.B.2 we know that the one-year forward price, V_{BBB} , of the five-year 6% coupon bond, if the obligor remains rated BBB, is

$$V_{BBB} = 6 + \frac{6}{1.041} + \frac{6}{1.0467^2} + \frac{6}{1.0525^3} + \frac{106}{1.0563^4} = 107.53$$

where the discount rates are taken from Table III.B.5.3. The cash-flows are shown in Figure III.B.5.4. If we replicate the same calculations for each rating category we obtain the values shown in Table III.B.5.4.⁴

Figure III.B.5.4: Cash flows for five-year 6% coupon bond



⁴ CreditMetrics calculates the forward value of the bonds, or loans, including compounded coupons paid out during the year.

Table III.B.5.4: One-year forward values for a BBB bond

Year-end rating	Value (\$)
AAA	109.35
AA	109.17
A	108.64
BBB	107.53
BB	102.00
B	98.08
CCC	83.62
Default	51.11

Source: CreditMetrics, J.P. Morgan

We do not assume that everything is lost if the issuer defaults at the end of the year. Depending on the seniority of the instrument, a recovery rate of par value is recuperated by the investor. These recovery rates are estimated from historical data by the rating agencies. Table III.B.5.5 shows the expected recovery rates for bonds by different seniority classes as estimated by Moody's.⁵ In simulations performed to assess the portfolio distribution, the recovery rates are not taken as fixed, but rather are drawn from a distribution of possible recovery rates. The distribution of the changes in the bond value, at the one-year horizon, due to an eventual change in credit quality is shown in Table III.B.5.6 and Figure III.B.5.5.

Table III.B.5.5: Recovery rates by seniority class (% of face value, i.e. 'par')

Seniority Class	Mean (%)	Standard Deviation (%)
Senior Secured	53.80	26.86
Senior Unsecured	51.13	25.45
Senior Subordinated	38.52	23.81
Subordinated	32.74	20.18
Junior Subordinated	17.09	10.90

Source: Carty and Lieberman (1996).

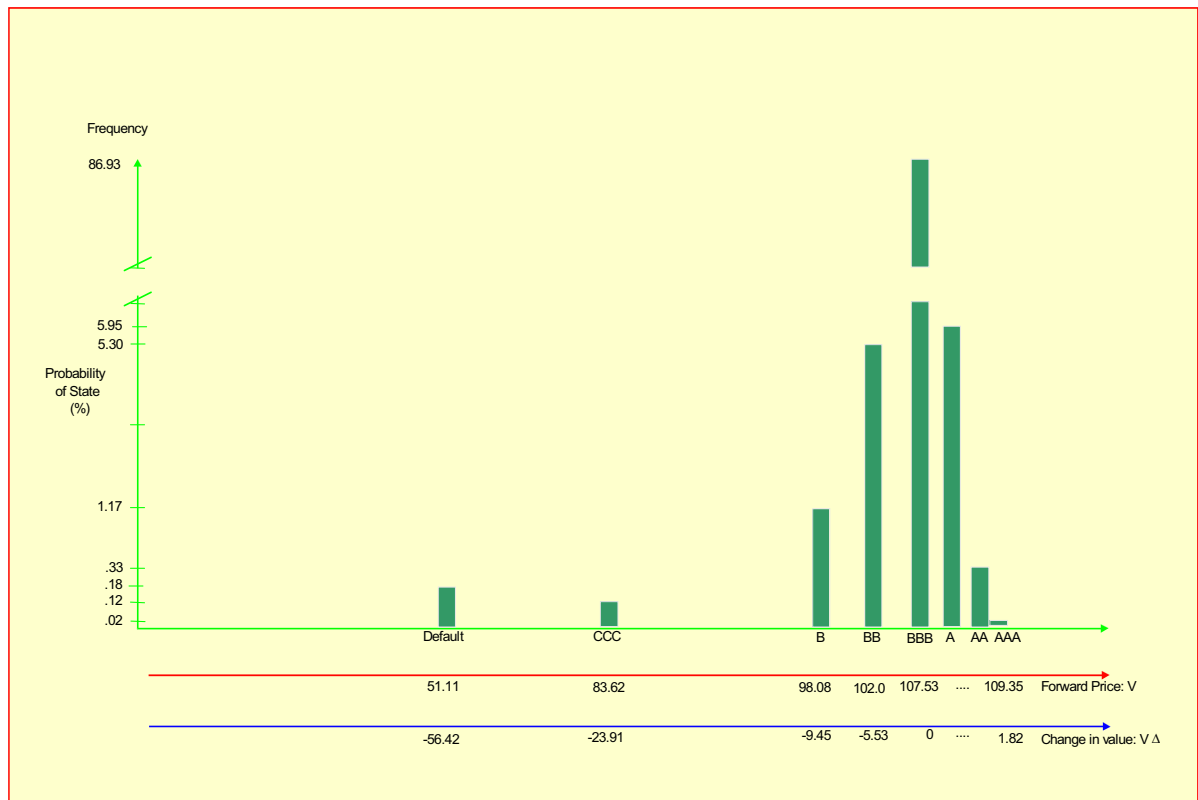
⁵ Cf. Carty and Lieberman (1996). See also Altman and Kishore (1996, 1998) for similar statistics.

Table III.B.5.6: Distribution of the bond values, and changes in value of a BBB bond, in 1 year

Year-end rating	Probability of state: $p(\%)$	Forward price: V (\$)	Change in value: ΔV (\$)
AAA	0.02	109.35	1.82
AA	0.33	109.17	1.64
A	5.95	108.64	1.11
BBB	86.93	107.53	0
BB	5.30	102.00	-5.53
B	1.17	98.08	-9.45
CCC	0.12	83.62	-23.91
Default	0.18	51.11	-56.42

Source: CreditMetrics, J.P. Morgan.

Figure III.B.5.5: Histogram of the one-year forward prices and changes in value of a BBB bond



This distribution exhibits a long ‘downside tail.’ The first percentile of the distribution of ΔV , which corresponds to credit VaR at the 99% confidence level, is -23.91 . It is a much lower value

than if we computed the first percentile assuming a normal distribution for ΔV . In that case credit VaR at the 99% confidence level would be only -7.43 .⁶

The above analysis is the basis for the evaluation of the portfolio of loans, along the lines described above. CreditMetrics proposes to use Monte Carlo simulations to assess the credit risk of a bond/loan portfolio due to the large amount of factors that must be taken into consideration.

III.B.5.3.2 Estimation of Default and Rating Changes Correlations

So far we have shown how the future distribution of values for a given bond (or loan) is derived. In what follows we focus on how to estimate the potential changes in the value of a portfolio of creditors, when the changes are due to credit risk only, and credit risk is expressed as the potential ratings changes during the year. One important factor in the portfolio assessment is the correlation between changes in the credit ratings and the default correlation for any two obligors. In reality, the correlations between the changes in credit quality are not zero, and the overall credit VaR is quite sensitive to these correlations. Their accurate estimation is therefore one of the key determinants of portfolio optimisation.

Default correlations might be expected to be higher for firms within the same industry, or in the same region, than for firms in unrelated sectors. In addition, correlations vary with the relative state of the economy in the business cycle. If there is a slowdown in the economy, or a recession, most of the assets of the obligors will decline in value and quality, and the likelihood of multiple defaults increases substantially. The opposite happens when the economy is performing well: default correlations go down. Thus, we cannot expect default and migration probabilities to remain stationary over time. There is clearly a need for a structural model that relates changes in default probabilities to fundamental variables. CreditMetrics derives the default and migration probabilities from a correlation model of the firm's assets.

CreditMetrics makes use of the stock price of a firm as a proxy for its asset value, as the true asset value is not directly observable. (This is another simplifying assumption in CreditMetrics

⁶ The mean, μ , and the variance, σ^2 , of the distribution for ΔV can be calculated from the data in Table III.B.5.5 as follows:

$$\begin{aligned}\mu &= \text{mean}(\Delta V) = \sum_i p_i \Delta V_i \\ &= 0.02\% \times 1.82 + 0.33\% \times 1.64 + \dots + 0.18\% \times (-56.42) \\ &= -0.46 \\ \sigma^2 &= \text{variance}(\Delta V) = \sum_i p_i (\Delta V_i - \mu)^2 \\ &= 0.02\% (1.82 - 0.46)^2 + 0.33\% (1.64 - 0.46)^2 + \dots + 0.18\% (-56.42 - 0.46)^2 = 8.95\end{aligned}$$

i.e. $\sigma = 2.99$. The 0.01 percentile of a normal distribution $N(\mu, \sigma^2)$ is $\mu - 2.33\sigma$, i.e. -7.43 .

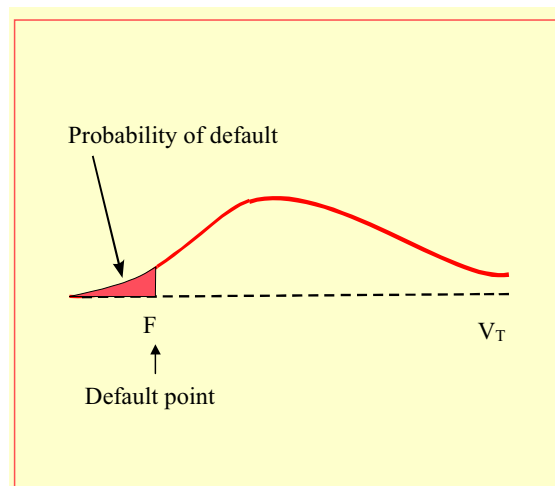
that may affect the accuracy of the approach.) CreditMetrics estimates the correlations between the equity returns of various obligors, then it infers the correlations between changes in credit quality directly from the joint distribution of these equity returns.

The theoretical framework underlying all this is the option pricing approach to the valuation of corporate securities first developed by Merton (1974). The model is described in detailed in Section III.B.5.5 as it forms the basis for the KMV approach. In Merton’s model, the firm is assumed to have a very simple capital structure; it is financed by equity, S_t , and a single zero-coupon debt instrument maturing at time T , with face value F , and current market value B_t . The firm’s balance sheet is represented in Table III.B.5.7, where V_t is the value of all the assets and $V_t = B_t(F) + S_t$.

Table III.B.5.7: Balance sheet of Merton’s firm

	Assets	Liabilities / Equity
	Risky Assets: V_t	Debt: $B_t(F)$
		Equity: S_t
Total:	V_t	V_t

Figure III.B.5.6: Distribution of the firm’s assets value at maturity of the debt obligation



In this framework, default occurs at the maturity of the debt obligation only when the value of assets is less than the payment, F , promised to the bondholders. Figure III.B.5.6 shows the distribution of the assets’ value at time T , the maturity of the zero-coupon debt, and the probability of default (i.e. the shaded area on the left-hand side of the default point, F).

Merton’s model is extended by CreditMetrics to include changes in credit quality as illustrated in Figure III.B.5.7. This generalisation consists of slicing the distribution of asset returns into bands in such a way that, if we draw randomly from this distribution, we reproduce exactly the migration frequencies as shown in the transition matrices that we discussed earlier.

Figure III.B.5.7 shows the distribution of the normalised assets’ rates of return, one year ahead. The distribution is normal with mean zero and unit variance. The credit rating ‘thresholds’ are calculated using the transition probabilities in Table III.B.5.1 for a BB-rated obligor. The area in the right-hand tail of the distribution, down to Z_{AAA} , corresponds to the probability that the obligor will be upgraded from BB to AAA, i.e. 0.03%. Then, the area between Z_{AA} and Z_{AAA} corresponds to the probability of being upgraded from BB to AA, etc. The area in the left-hand tail of the distribution, to the left of Z_{CCG} , corresponds to the probability of default, i.e. 1.06%.

Figure III.B.5.7: Generalisation of the Merton model to include rating changes

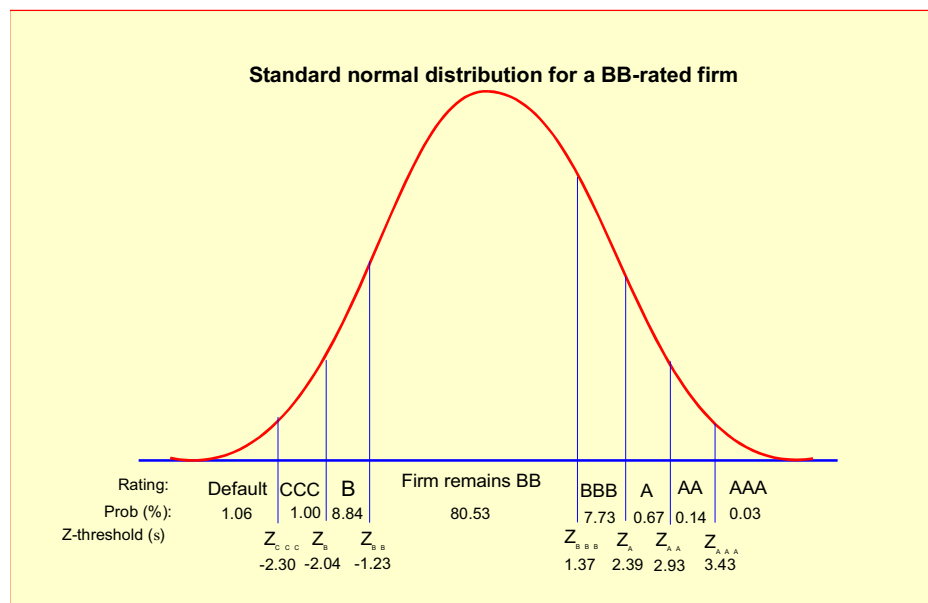


Table III.B.5.8: Transition probabilities and credit quality thresholds for BB- and A-rated obligors

Rating in one year	A-rated obligor		BB-rated obligor	
	Probabilities (%)	Thresholds: Z (σ)	Probabilities (%)	Thresholds: Z (σ)
AAA	0.09	3.12	0.03	3.43
AA	2.27	1.98	0.14	2.93
A	91.05	-1.51	0.67	2.39
BBB	5.52	-2.30	7.73	1.37
BB	0.74	-2.72	80.53	-1.23
B	0.26	-3.19	8.84	-2.04
CCC	0.01	-3.24	1.00	-2.30
Default	0.06		1.06	

Table III.B.5.8 shows the transition probabilities for two obligors rated BB and A respectively, and the corresponding credit quality thresholds. The thresholds are given in terms of normalised standard deviations. For example, for a BB-rated obligor the default threshold is -2.30 standard deviations from the mean rate of return.

This generalisation of Merton's model is quite easy to implement. It assumes that the normalised log-returns over any period of time are normally distributed with a mean of 0 and a variance of 1, and the distribution is the same for all obligors within the same rating category. If p_{Def} denotes the probability of the BB-rated obligor defaulting, then the critical asset value V_{Def} is such that:

$$p_{Def} = \Pr(V_t \leq V_{Def})$$

which can be translated into a normalised threshold Z_{CCC} , such that the area in the left-hand tail below Z_{CCC} is p_{Def} .⁷ Z_{CCC} is simply the threshold point in the standard normal distribution, $N(0,1)$, corresponding to a cumulative probability of p_{Def} . Then, based on the option pricing model, the critical asset value V_{Def} which triggers default is such that $Z_{CCC} = -d_2$. This critical asset value

V_{Def} is also called the *default point*.⁸

⁷ See the Appendix for the derivation of the proof. In the next section we define the 'distance to default' as the distance between the expected asset value and the default point.

⁸ Note that d_2 is different from its equivalent in the Black-Scholes formula since, here, we work with the 'actual' instead of the 'risk-neutral' return distributions, so that the drift term in d_2 is the expected return on the firm's assets, instead of the risk-free interest rate as in Black-Scholes. See Chapter I.A.8 for the definition of d_2 in Black-Scholes and Appendix 1 for d_2 in the above derivation.

Note that only the threshold levels are necessary to derive the joint migration probabilities, and these can be calculated without it being necessary to observe the asset value, and to estimate its mean and variance. To derive the critical asset value V_{Def} we only need to estimate the expected asset return μ and asset volatility σ . Accordingly Z_B is the threshold point corresponding to a cumulative probability of being either in default or in rating CCC, i.e. $p_{Def} + p_{CCG}$, etc.

We mentioned above that, as asset returns are not directly observable, CreditMetrics makes use of equity returns as their proxy. Yet using equity returns in this way is equivalent to assuming that all the firm's activities are financed by means of equity. This is a major drawback of the approach, especially when it is being applied to highly leveraged companies. For those companies, equity returns are substantially more volatile, and possibly less stationary, than the volatility of the firm's assets.

Now, assume that the correlation between the assets' rates of return is known, and is denoted by ρ , which is assumed to be equal to 0.2 in our example. The normalised log-returns on both assets follow a joint normal distribution:

$$f(r_{BB}, r_A; \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{ \frac{-1}{2(1-\rho^2)} [r_{BB}^2 - 2\rho r_{BB}r_A + r_A^2] \right\}$$

We can therefore compute the probability of both obligors being in any particular combination of ratings. For example, we can compute the probability that they will remain in the same rating classes, i.e. BB and A, respectively:

$$\Pr(-1.23 < r_{BB} < 1.37, -1.51 < r_A < 1.98) = 0.7365$$

where r_{BB} and r_A are the rates of return on the assets of obligors BB and A, respectively, assumed normally distributed as in Figure III.B.5.7.⁹

For any two obligors the joint probability of both obligors defaulting is

$$p_{1,2} = \Pr[V_1 \leq V_{Def1}, V_2 \leq V_{Def2}]$$

where V_1 and V_2 and denote the asset values for both obligors at time t , and V_{Def1} and V_{Def2} are the corresponding default points. This joint probability may be calculated in exactly the same way as the migration probabilities were calculated above, i.e. using the bivariate normal distribution.

⁹ See Chapter II.E for details on how to compute joint probabilities when the two random variables have a bivariate normal distribution.

Given this joint probability of default, and the individual probabilities of default for each obligor, p_1 and p_2 , the default correlation can be calculated as:¹⁰

$$\text{corr}(Def1, Def2) = \frac{p_{1,2} - p_1 p_2}{\sqrt{p_1(1-p_1)p_2(1-p_2)}} \quad (\text{III.B.5.1})$$

We can illustrate the results with a numerical example. If the probabilities of default for obligors rated A and BB are $P_{Def}(A) = 0.0006$ and $P_{Def}(BB) = 0.0106$, respectively, and the correlation coefficient between the rates of return on the two assets is $\rho = 0.2$, then the joint probability of default is only 0.000054.¹¹ Now, using equation (III.B.5.1), we find that the correlation coefficient between the two default events is only 0.019.

This example, with asset return correlation of 0.2 but default correlation of only 0.019, is not unusual. Asset returns correlations are approximately 10 times larger than default correlations for asset correlations in the range of 0.2–0.6%. This shows that the joint probability of default is in fact quite sensitive to pairwise asset return correlations, and it illustrates how important it is to estimate these data correctly if one is to assess the diversification effect within a portfolio. It can be shown that the impact of correlations on credit VaR is quite large. It is larger for portfolios with relatively low-grade credit quality than it is for high-grade portfolios. Indeed, as the credit quality of the portfolio deteriorates and the expected number of defaults increases, this number is magnified by an increase in default correlations.

III.B.5.3.3 Credit VaR of a Bond/Loan Portfolio

The analytic approach that we sketched out above for a portfolio with bonds issued by two obligors is not practicable for large portfolios. Instead, CreditMetrics implements a Monte Carlo simulation to generate the full distribution of the portfolio values at the credit horizon of one year. The following steps are necessary:

1. Derive the asset return thresholds for each rating category.
2. Estimate the correlation between each pair of obligors' asset returns.
3. Generate return scenarios according to their joint normal distribution. A standard technique that is often used to generate correlated normal variables is the Cholesky decomposition.¹² Each scenario is characterised by n standardised asset returns, one for each of the n obligors in the portfolio.

¹⁰ See Lucas (1995).

¹¹ If the default events were independent the joint probability of default would simply be the product of the two default probabilities, i.e. $0.0006 \times 0.0106 = 0.000064$.

¹² A good reference on Monte Carlo simulations and the Cholesky decomposition is Fishman (1997, p. 223)

4. For each scenario, and for each obligor, map the standardised asset return into the corresponding rating, according to the threshold levels derived in step 1.
5. Given the spread curves, which apply for each rating, revalue the portfolio.
6. Repeat the procedure a large number of times, say 100,000, and plot the distribution of the portfolio values to obtain a graph such as Figure III.B.5.1.
7. Finally, derive the percentiles of the distribution of the future values of the portfolio to obtain the credit VaR and/or credit economic capital as in Figure II.B.5.2.

Estimating VaR for credit requires a very large number of simulations as the loss distribution is very skewed with very few observations in the tail. In order to reduce substantially the number of simulations, say by a factor of 5–10, while maintaining the same level of accuracy, it is recommended for practical applications to implement credit portfolio models with the use of variance reduction techniques. ‘Importance sampling’ is a technique which produces remarkable results for credit risk (Glasserman *et al.*, 2000).

III.B.5.4 Conditional Transition Probabilities– CreditPortfolioView

CreditPortfolioView is a multi-factor model that is used to simulate the joint conditional distribution of default and migration probabilities for various rating groups in different industries, and for each country, conditional on the value of macro-economic factors. CreditPortfolioView is based on the observation that default probabilities and credit migration probabilities are linked to the economy. When the economy worsens both downgrades and defaults increase; when the economy becomes stronger, the contrary holds true. In other words, credit cycles follow business cycles closely.

Since the shape of the economy is, to a large extent, driven by macro-economic factors, CreditPortfolioView proposes a methodology to link those macro-economic factors to default and migration probabilities. It employs the values of macro-economic factors such as the unemployment rate, the rate of growth in GDP, the level of long-term interest rates, foreign exchange rates, government expenditures and the aggregate savings rate.

Provided that data are available, this methodology can be applied in each country to various sectors and various classes of obligors that react differently during the business cycle – sectors such as construction, financial institutions, agriculture, and services. It applies better to speculative-grade obligors whose default probabilities vary substantially with the credit cycle, than to investment-grade obligors whose default probabilities are more stable.

Conditional default probabilities are modelled as a *logit function*, whereby the independent variable is a country-specific index that depends upon current and lagged macro-economic variables. That is:¹³

$$P_{j,t} = \frac{1}{1 + e^{-Y_{j,t}}}$$

where $P_{j,t}$ is the conditional probability of default in period t , for speculative-grade obligors in country/industry j , and $Y_{j,t}$ is the country index value derived from a multi-factor model.¹⁴

In order to derive the conditional transition matrix the (unconditional) transition matrix based on Moody's or Standard & Poor's historical data will be used. These transition probabilities are unconditional in the sense that they are historical averages based on more than 20 years of data covering several business cycles, across many different countries and industries. As we discussed earlier, default probabilities for non-investment grade obligors are higher than average during a period of recession. Also credit downgrades increase, while upward migrations decrease. The opposite holds during a period of economic expansion. We can express this in the following way:

$$\begin{aligned} \frac{P_{j,t}}{\varphi P_{j,t}} &> 1 \text{ in economic recession} \\ \frac{P_{j,t}}{\varphi P_{j,t}} &< 1 \text{ in economic expansion} \end{aligned} \tag{III.B.5.2}$$

where $\varphi P_{j,t}$ is the unconditional (historical average) probability of default in period t , for speculative-grade obligors in country/industry j . CreditPortfolioView proposes to use (III.B.5.2) to adjust the unconditional transition probabilities in order to produce a transition matrix \mathbf{M}_t that is conditional on the state of the economy:

$$\mathbf{M}_t = \mathbf{M}(P_{j,t}/\varphi P_{j,t})$$

where the adjustment consists in shifting the probability mass toward downgraded and defaulted states when the ratio $P_{j,t}/\varphi P_{j,t}$ is greater than one, and in the opposite direction if the ratio is less than one. Since one can simulate $P_{j,t}$ over any time horizon $t = 1, \dots, T$, this approach can generate multi-period transition matrices:

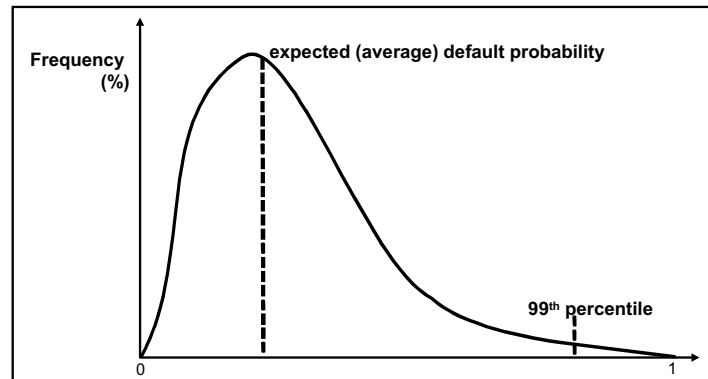
$$\mathbf{M}_T = \prod_{t=1, \dots, T} \mathbf{M}(P_{j,t} / \varphi P_{j,t}). \tag{III.B.5.3}$$

¹³ Note that the logit function ensures that the probability takes a value between 0 and 1.

¹⁴ J.P. Morgan (1997) provides an example of multi-factor model in the context of credit risk modelling. For a review of multifactor models, see Elton and Gruber (1995) and Rudd and Clasing (1988).

One can simulate the transition matrix (III.B.5.3) many times to generate a distribution of the cumulative conditional default probability such as that shown in Figure III.B.5.8 for any rating over any time period. The same Monte Carlo methodology can be used to produce the conditional cumulative distributions of migration probabilities over any time horizon.

Figure III.B.5.8: Distribution of the cumulative conditional default probability, for a given rating, over a given time horizon, T



CreditPortfolioView and KMV (described in Section III.B.5.6) base their approach on the empirical observation that default and migration probabilities vary over time. KMV adopts a micro-economic approach that relates the probability of default of any obligor to the market value of its assets. CreditPortfolioView proposes a methodology that links macro-economic factors to default and migration probabilities. The calibration of CreditPortfolioView thus requires reliable default data for each country, and possibly for each industry sector within each country.

Another limitation of the model is the *ad hoc* adjustment of the transition matrix. It is not clear that the proposed methodology performs better than a simple Bayesian model, where the revision of the transition probabilities would be based on the internal expertise accumulated by the credit department of the bank, and an internal appreciation of the current stage of the credit cycle (given the quality of the bank's credit portfolio). These two approaches are somewhat related since the market value of the firms' assets depends on the shape of the economy; it would be interesting to compare the transition matrices produced by both models.

III.B.5.5 The Contingent Claim Approach to Measuring Credit Risk

The CreditMetrics approach to measuring credit risk, as described previously, is rather appealing as a methodology. Unfortunately it has a major weakness: reliance on ratings transition

probabilities that are based on average historical frequencies of defaults and credit migration. As a result, the accuracy of CreditMetrics calculations depends upon two critical assumptions: first, that all firms within the same rating class have the same default rate and the same spread curve, even when recovery rates differ among obligors; and second, that the actual default rate is equal to the historical average default rate. Credit rating changes and credit quality changes are taken to be identical. Credit rating and default rates are also synonymous, i.e. the rating changes when the default rate is adjusted, and vice versa. This view has been strongly challenged by researchers working for the consulting and software corporation KMV.¹⁵ Indeed, the assumption cannot be true since we know that default rates evolve continuously, while ratings are adjusted in a discrete fashion. (This lag is because rating agencies necessarily take time to upgrade or downgrade companies whose default risk has changed.)

What we call the ‘structural’ approach offers an alternative to the credit migration approach. Here, the economic value of default is presented as a put option on the value of the firm’s assets. The merit of this approach is that each firm can be analysed individually based on its unique features. But this is also the principal drawback, since the information required for such an analysis is rarely available to the bank or the investor. One has to estimate the total value, and the risk (e.g. the volatility) of the firm’s assets. The option pricing approach, introduced by Merton (1974) in a seminal paper, builds on the limited liability rule which allows shareholders to default on their obligations while they surrender the firm’s assets to the various stakeholders, according to pre-specified priority rules. The firm’s liabilities are thus viewed as contingent claims issued against the firm’s assets, with the payoffs to the various debt-holders completely specified by seniority and safety covenants. Default occurs at debt maturity whenever the firm’s asset value falls short of debt value at that time. In this model, the loss rate is endogenously determined and depends on the firm’s asset value, volatility, and the default-free interest rate for the debt maturity.

III.B.5.5.1 Structural Model of Default Risk: Merton’s (1974) Model

To determine the value of the credit risk arising from a bank loan, we first make two assumptions: that the loan is the only debt instrument of the firm, and that the only other source of financing is equity. In this case, as we shall see below, the credit value is equal to the value of a put option on the value of assets of the firm, at a strike price equal to the face value of debt (including accrued interest), maturing at the maturity of the debt. By purchasing the put on the assets of the firm for the term of the debt, with a strike price equal to the face value of the loan, the bank can completely eliminate all the credit risk and convert the risky corporate loan into a

¹⁵ KMV is a trademark of KMV Corporation. The initials KMV stand for the surnames of Stephen Kealhofer, John McQuown and Oldrich Vasicek who founded KMV Corporation in 1989. Kealhofer and Vasicek are former academics from the University of California at Berkeley.

riskless loan. Thus, the cost of eliminating the credit risk associated with providing a loan to the firm is the value of this put option. Now, if we make the assumptions that are needed to apply the Black–Scholes (BS) model (Black and Scholes, 1973) to equity and debt instruments, we can express the value of the credit risk in an option-like formula.

Consider a firm with risky assets V , which is financed by equity, S , and by one debt obligation, maturing at time T with face value (including accrued interest) of F and market value B . If we assume that markets are frictionless, with no taxes, and there is no bankruptcy cost, then the value of the firm’s assets is simply the sum of the firm’s equity and debt. At time $t = 0$ then,

$$V_0 = S_0 + B_0. \quad (\text{III.B.5.4})$$

The loan to the firm is subject to credit risk, namely the risk that at time T the value of the firm’s assets V_T , will be below the obligation to the debt holders, F . Credit risk exists as long as the probability of default, $\Pr(V_T < F)$, is greater than zero.

From the viewpoint of a bank that makes a loan to the firm, this gives rise to a series of questions. Can the bank eliminate/reduce credit risk, and at what price? What is the economic cost of reducing credit risk? And, what are the factors affecting this cost? In this simple framework, credit risk is a function of the financial structure of the firm, i.e.

- its *leverage ratio* $L \equiv Fe^{rT}/V_0$, where V_0 is the present value of the firm’s assets, and Fe^{-rT} is the present value of the debt obligation at maturity,
- the *volatility* σ of the firm’s assets, and
- the time T to *maturity of the debt*.

The model was initially suggested by Merton (1974) and further analysed by Galai and Masulis (1976). To understand why the credit value is equal to the value of a put option on the value of assets of the firm, at a strike price equal to the face value of debt (including accrued interest), and with maturity equal to the maturing of the debt, consider Table III.B.5.9. This shows that if the bank buys the put option with value P , the value at time T will be F whether $V_T \leq F$ or $V_T > F$, so credit risk is eliminated. In this way they can convert the risky corporate loan into a riskless loan with a face value of F .

Table III.B.5.9: Bank’s payoff at times 0 and T for making a loan and buying a put option

Time	0	T	
Value of assets	V_0	$V_T \leq F$	$V_T > F$
Bank’s position:			
(a) make a loan	$-B_0$	V_T	F
(b) buy a put	$-P_0$	$F - V_T$	0
Total	$-B_0 - P_0$	F	F

Thus the value of the put option is the cost of eliminating the credit risk associated with providing a loan to the firm. If we make the assumptions that are needed to apply the (BS) model to equity and debt instruments (see Galai and Masulis, 1976, for a detailed discussion of the assumptions), we can write the value of the put as:

$$P_0 = -N(-d_1)V_0 + Fe^{-rT}N(-d_2), \quad (\text{III.B.5.5})$$

where P_0 is the current value of the put, $N(\cdot)$ is the cumulative standard normal distribution,

$$d_1 = \frac{\ln(V_0 / F) + (r + \sigma^2 / 2)T}{\sigma\sqrt{T}} = \frac{\ln(V_0 / Fe^{-rT}) + \sigma^2 T / 2}{\sigma\sqrt{T}}, \quad d_2 = d_1 - \sigma\sqrt{T}, \quad (\text{III.B.5.6})$$

and σ is the standard deviation of the rate of return of the firm’s assets. The model illustrates that the credit risk, and its costs, is an increasing function of the volatility of the assets of the firm σ and the time interval T until debt is paid back, and a decreasing function of the risk-free interest rate r (the higher is r , the less costly it is to reduce credit risk). The cost of credit risk is also homogeneous function of the leverage ratio, which means that it stays constant for a scale expansion of Fe^{-rT} / V_0 .

Note that, when the probability of default is greater than zero the yield to maturity on the debt y_T must be greater than the risk-free rate r , so that the *default spread* $\pi_T = y_T - r$ that compensates the bond holders for the default risk that they bear is positive. The default spread can be regarded as a risk premium associated with holding risky bonds. It can be shown that, in the Merton (1974) framework, the default spread can be computed exactly as a function of the leverage ratio, the volatility of the underlying assets and the debt maturity. In fact:

$$\pi_T = y_T - r = -\frac{1}{T} \ln \left(N(d_2) + \frac{V_0}{Fe^{-rT}} N(-d_1) \right).$$

Note that the default spread decreases when the risk-free rate increases. The greater the risk-free rate, the less risky is the bond and the lower is the value of the put protection – therefore, the lower is the risk premium. The numerical examples in Table III.B.5.10 show the default spread for various levels of volatility and different leverage ratios.

Table III.B.5.10: Default spread for corporate debt

(for $V_0 = 100$, $T = 1$, and $r = 10\%$ ¹⁶)

Leverage ratio: L	Volatility of underlying asset: σ			
	0.05	0.10	0.20	0.40
0.5	0	0	0	1.0%
0.6	0	0	0.1%	2.5%
0.7	0	0	0.4%	5.6%
0.8	0	0.1%	1.5%	8.4%
0.9	0.1%	0.8%	4.1%	12.5%
1.0	2.1%	3.1%	8.3%	17.3%

Example III.B.5.1

We show how the 5.6% default spread (marked in red) was obtained in Table III.B.5.10. Using equations (III.B.5.5) and (III.B.5.6) with $V_0 = 100$, $T = 1$, $r = 0.1$ (i.e. 10%), $\sigma = 0.4$ (i.e. 40%) with the leverage ratio $L = 70\%$, we obtain $S_0 = 33.37$ for the value of equity and $B_0 = 66.63$ for the value of the corporate risky debt. Since $L = Fe^{-rT} / V_0$, we have $F = 77$. Therefore the yield on the loan is $77/66.63 - 1 = 0.156$ and there is a 5.6% risk premium to reflect the credit risk.

The model also shows that the put value is $P_0 = 3.37$. Hence the cost of eliminating the credit risk is \$3.37 for \$100 worth of the firm’s assets, where the face value (i.e. the principal amount plus the promised interest rate) of the one-year debt is 77. This cost drops to 25 cents when volatility decreases to 20% and to 0 for 10% volatility. The assets’ volatility is clearly a critical factor in determining credit risk. To demonstrate that the bank eliminates all its credit risk by buying the put, we can compute the yield on the bank’s position as

$$F / (B_0 + P) = 77 / (66.63 + 3.37) = 1.10,$$

which translates to a riskless yield of 10% per annum.

III.B.5.5.2 Estimating Credit Risk as a Function of Equity Value

We have already shown that the cost of eliminating credit risk can be derived from the value of the firm’s assets. A practical problem arises over how easy it is to observe V . In some cases, if both equity and debt are traded, V can be reconstructed by adding the market values of both equity and debt. However, corporate loans are not often traded and so, to all intents and

¹⁶ 10% is the annualised interest rate discreetly compounded, which is equivalent to 9.5% continuously compounded.

purposes, we can only observe equity. The question, then, is whether the risk of default can be hedged by trading shares and derivatives on the firm's stock.

In the Merton framework, equity itself is a contingent claim on the firm's assets. Its value can be expressed as a function of the same parameters as the put option:

$$S = VN(d_1) - Fe^{-rT}N(d_2) \quad (\text{III.B.5.7})$$

A put can be created synthetically by selling short $N(-d_1)$ units of the firm's assets, and buying F $N(-d_2)$ units of government bonds maturing at T , with face value of F . If one sells short $N(-d_1)/N(d_1)$ units of the stock S , one effectively creates a short position in the firm's assets of $N(-d_1)$ units, since:

$$\frac{-N(-d_1)}{N(d_1)}S = -VN(-d_1) + Fe^{-rT}N(d_2)\frac{N(-d_1)}{N(d_1)}.$$

Therefore, if V is not directly traded or observed, one can create a put option dynamically by selling short the appropriate number of shares. The equivalence between the put and the synthetic put is valid over short time intervals, and must be readjusted frequently with changes in S and in time left to debt maturity.

Example III.B.5.2

Using the data from Example III.B.5.1, $N(-d_1)/N(d_1) = -0.137/0.863 = -0.159$. This means that in order to insure against the default of a one-year loan with a maturity value of 77, for a firm with a current market value of assets of 100, the bank should sell short 0.159 of the outstanding equity. Note that the outstanding equity is equivalent to a short-term holding of $N(d_1) = 0.863$ of the firm's assets. Shorting 0.159 of equity is equivalent to shorting 0.863 of the firm's assets.

The question now is whether we can use a put option on equity in order to hedge the default risk. It should be remembered that equity itself reflects the default risk, and as a contingent claim its instantaneous volatility σ_S can be expressed as:

$$\sigma_S = \eta_{S,V}\sigma \quad (\text{III.B.5.8})$$

where $\eta_{S,V} = N(d_1)V/S$ is the instantaneous elasticity of equity with respect to the firm's value, and $\eta_{S,V} \geq 1$. Since σ_S is stochastic and changes with V , the conventional BS model cannot be applied to the valuation of puts and call on S . The BS model requires σ to be constant, or to follow a deterministic path over the life of the option. However in practice, for long-term options, the estimated σ_S from (III.B.5.8) is not expected to change widely from day to day.

Therefore, equation (III.B.5.8) can be used in the context of BS estimation of long-term options, even when the underlying instrument does not follow a stationary lognormal distribution.

III.B.5.6 The KMV Approach

KMV derives the *expected default frequency* (EDF), i.e. the default probability, for each obligor based on the Merton (1974) type of model. The probability of default is thus a function of the firm's capital structure, the volatility of the asset returns and the current asset value. The EDF is firm-specific, and can be mapped onto any rating system to derive the equivalent rating of the obligor. EDFs can be viewed as a 'cardinal ranking' of obligors relative to default risk, instead of the more conventional 'ordinal ranking' proposed by rating agencies (which relies on letters such as AAA, AA, ...).

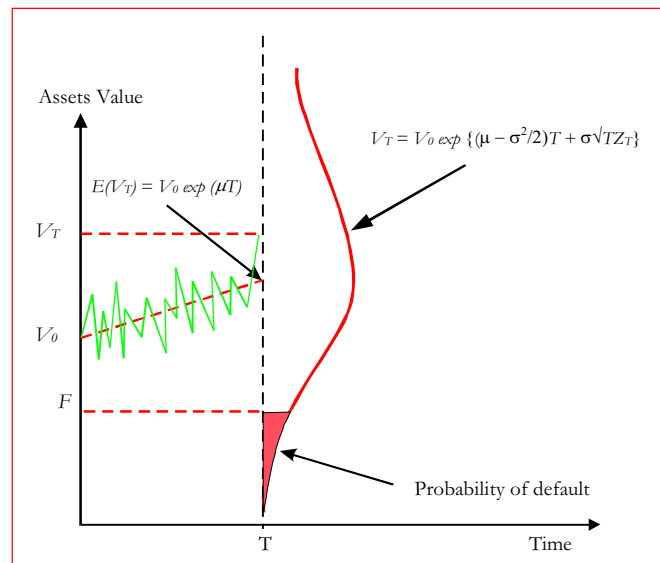
Contrary to CreditMetrics, KMV's model does not make any explicit reference to the transition probabilities which, in KMV's methodology, are already embedded in the EDFs. Indeed, each value of the EDF is associated with a spread curve and an implied credit rating.

Credit risk in the KMV approach is essentially driven by the dynamics of the asset value of the issuer. Given the capital structure of the firm,¹⁷ and once the stochastic process for the asset value has been specified, *the actual probability of default* for any time horizon, one year, two years, etc., can be derived. Figure III.B.5.9 depicts how the probability of default relates to the distribution of asset returns and the capital structure of the firm.

We assume that the firm has a very simple capital structure. It is financed by means of equity, S_t , and a single zero-coupon debt instrument maturing at time T , with face value F , and current market value B_t . The firm's balance sheet can be represented as follows: $V_t = B_t(F) + S_t$, where V_t is the value of all the assets. The value of the firm's assets, V_t , is assumed to follow a standard geometric Brownian motion. In this framework, default only occurs at maturity of the debt obligation, when the value of assets is less than the promised payment F to the bondholders. Figure III.B.5.9 shows the distribution of the assets' value at time T , the maturity of the zero-coupon debt, and the probability of default which is the shaded area below F .

¹⁷ That is, the composition of its liabilities: equity, short- and long-term debt, convertible bonds, etc.

Figure III.B.5.9: Distribution of the firm’s assets value at maturity of the debt obligation



The KMV approach is best applied to publicly traded companies, where the value of the equity is determined by the stock market. The information contained in the firm’s stock price and balance sheet can then be translated into an implied risk of default, as shown in the next section. The derivation of the actual probabilities of default proceeds in three stages:

- estimation of the market value and volatility of the firm’s assets;
- calculation of the distance to default, which is an index measure of default risk; and
- scaling of the distance to default to actual probabilities of default using a default database.

III.B.5.6.1 Estimation of the Asset Value V_A and the Volatility of Asset Return σ_A

In the contingent claim approach to the pricing of corporate securities, the market value of the firm’s assets is assumed to be lognormally distributed, i.e. the log-asset return follows a normal distribution.¹⁸ This assumption is quite robust and, according to KMV’s own empirical studies, actual data conform quite well to this hypothesis.¹⁹ In addition, the distribution of asset returns is stable over time, i.e. the volatility of asset returns remains relatively constant.

As we discussed earlier, if all the liabilities of the firm were traded, and marked to market every day, then the task of assessing the market value of the firm’s assets and its volatility would be

¹⁸ Financial models consider essentially market values of assets, and not accounting values, or book values, which only represent the historical cost of the physical assets, net of their depreciation. Only the market value is a good measure of the value of the firm’s ongoing business and it changes as market participants revise the firm’s future prospects. KMV models the market value of assets. In fact, there might be huge differences between both the market and the book values of total assets. For example, as of February 1998 KMV has estimated the market value of Microsoft assets at US\$228.6 billion versus \$16.8 billion for their book value, while for Trump Hotel and Casino the book value, which amounts to \$2.5 billion, is higher than the market value of \$ 1.8 billion.

¹⁹ The exception is when the firm’s portfolio of businesses has changed substantially through mergers and acquisitions, or restructuring.

straightforward. The firm's asset value would be simply the sum of the market values of the firm's liabilities, and the volatility of the asset return could be simply derived from the historical time series of the reconstituted asset value. In practice, however, only the price of equity for most public firms is directly observable, and in some cases part of the debt is actively traded.

The alternative approach to assets valuation consists of applying the option-pricing model to the valuation of corporate liabilities as suggested in Merton (1974). In order to make their model tractable KMV assume that the capital structure of a corporation is composed solely of equity, short-term debt (considered equivalent to cash), long-term debt (in perpetuity), and convertible preferred shares.²⁰ Given these simplifying assumptions, it is possible to derive analytical solutions for the value of equity, S , and its volatility, σ_S :

$$S = f(V, \sigma, L, \epsilon, r), \quad (\text{III.B.5.9})$$

$$\sigma_S = g(V, \sigma, L, \epsilon, r), \quad (\text{III.B.5.10})$$

where L denotes the leverage ratio in the capital structure, ϵ is the average coupon paid on the long-term debt and r is the risk-free interest rate.

If σ_S were directly observable, like the stock price, we could simultaneously solve (III.B.5.9) and (III.B.5.10) for V and σ . But the instantaneous equity volatility, σ_S , is relatively unstable, and is in fact quite sensitive to the change in asset value; there is no simple way to measure σ_S precisely from market data. Since only the value of equity S is directly observable, we can back out V from (III.B.5.9) so that it becomes a function of the observed equity value, or stock price, and the volatility of asset returns:

$$V = b(S, \sigma, L, \epsilon, r) \quad (\text{III.B.5.11})$$

Here volatility is an *implicit* function of V , S , L , ϵ and r . So to calibrate the model for σ , KMV uses an iterative technique.

III.B.5.6.2 Calculation of the 'Distance to Default'

Using a sample of several hundred companies, KMV observed that firms default when the asset value reaches a level that is somewhere between the value of total liabilities and the value of short-term debt. Therefore, the tail of the distribution of asset values below total debt value may not be an accurate measure of the actual probability of default. Loss of accuracy may also result from factors such as the non-normality of the asset return distribution, and the simplifying assumptions made about the capital structure of the firm. This may be further aggravated if a

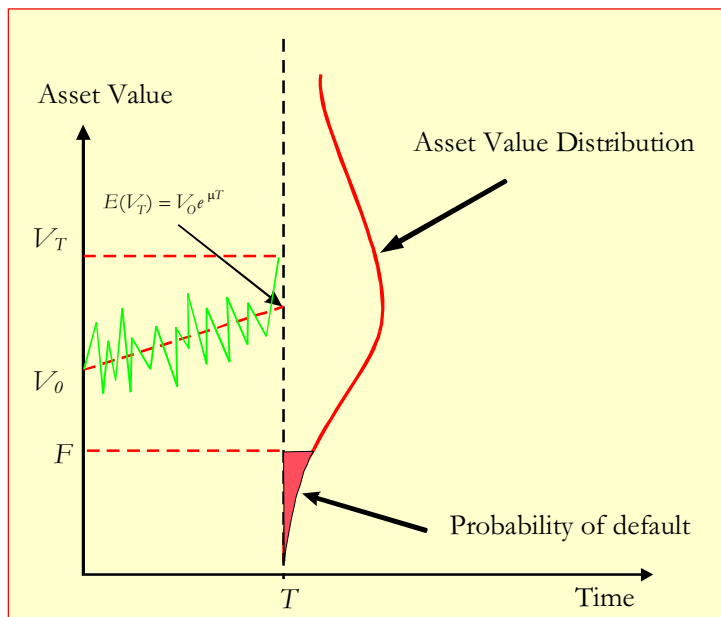
²⁰ In the general case the resolution of this model may require the implementation of complex numerical techniques, with no analytical solution, due to the complexity of the boundary conditions attached to the various liabilities. See, for example, Vasicek (1997).

company is able to draw on (otherwise unobservable) lines of credit. If the company is in distress, using these lines may (unexpectedly) increase its liabilities while providing the necessary cash to honour promised payments.

For all these reasons, KMV implements an intermediate phase before computing the probabilities of default. As shown in Figure III.B.5.10, which is similar to Figure III.B.5.9, KMV computes an index called *distance to default* (*DD*). This is the number of standard deviations between the mean of the distribution of the asset value, and a critical threshold called the ‘default point’ (*DPT*) which is set at the par value of current liabilities including short-term debt to be serviced over the time horizon (*STD*), plus half the long-term debt (*LTD*), i.e. $STD + LTD/2$. If the expected asset value in one year is $E(V_1)$ and σ is the standard deviation of future asset returns then

$$DD = \frac{E(V_1) - DPT}{\sigma}$$

Figure III.B.5.10: Distance to default



Given the lognormality assumption of asset values, the distance to default expressed in unit of asset return standard deviation at time horizon T , is

$$DD = \frac{\ln(V_0 / DPT_T) + (\mu - \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}} \quad (III.B.5.12)$$

where

- V_0 = current market value of assets
- DPT_T = default point at time horizon T
- μ = expected return on assets, net of cash outflows

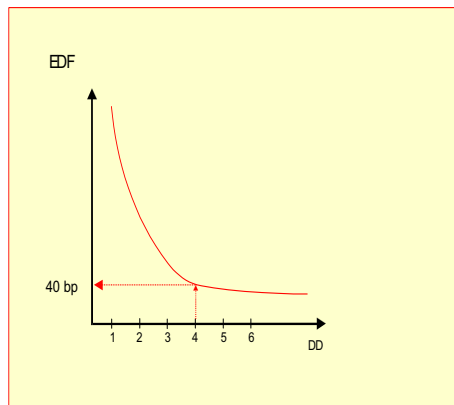
σ = annualised asset volatility.

It follows that the shaded area shown below the default point in Figure III.B.5.10 is equal to $N(-DD)$.

III.B.5.6.3 Derivation of the Probabilities of Default from the Distance to Default

This last phase consists of mapping the distance to default to the actual probabilities of default, for a given time horizon. KMV calls these probabilities *expected default frequencies*. Using historical information about a large sample of firms, including firms that have defaulted, one can estimate, for each time horizon, the proportion of firms of a given ranking, say $DD = 4$, that actually defaulted after one year. This proportion, say 40 bp, or 0.4%, is the EDF as shown in Figure III.B.5.11.

Figure III.B.5.11: Mapping of the ‘distance to default’ into the EDFs, for a given time horizon



Example III.B.5.3:

Current market value of assets:	$V_0 = 1000$
Net expected growth of assets per annum:	20%
Expected asset value in one year:	$V_0 \times 1.20 = 1200$
Annualised asset volatility, σ :	100
Default point:	800

Then $DD = (1200 - 800)/100 = 4$. Assume that among the population of 5000 firms with a DD of 4 at one point in time, 20 defaulted one year later. Then $EDF_{1 \text{ year}} = 20/5000 = 0.04 = 0.4\%$ or 40 bp. The implied rating for this probability of default is BB+.

Example III.B.5.4: Federal Express (\$ figures are in billions of US\$)

This example is provided by KMV and relates to Federal Express on two different dates: November 1997 and February 1998.

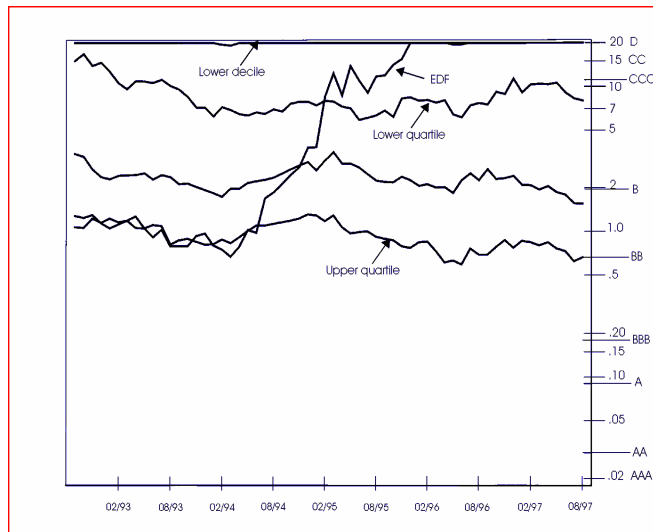
	November 1997	February 1998
Market capitalisation (price × shares outstanding)	\$ 7.7	\$ 7.3
Book liabilities	\$ 4.7	\$ 4.9
Market value of assets	\$ 12.6	\$ 12.2
Asset volatility	15%	17%
Default point	\$ 3.4	\$ 3.5
Distance to default (DD)	$\frac{12.6 - 3.4}{0.15 \times 12.6} = 4.9$	$\frac{12.2 - 3.5}{0.17 \times 12.2} = 4.2$
EDF	0.06% (6 bp) ≡ AA–	0.11% (11 bp) ≡ A–

This example illustrates the main causes of changes for an EDF, i.e. variations in the stock price, the debt level (leverage ratio), and asset volatility (i.e. the perceived degree of uncertainty concerning the value of the business).

III.B.5.6.4 EDF as a Predictor of Default

KMV has provided a ‘Credit Monitor’ service for estimated EDFs since 1993. EDFs have proved to be a useful leading indicator of default, or at least of the degradation of the creditworthiness of issuers. When the financial situation of a company starts to deteriorate, EDFs tend to shoot up quickly until default occurs, as shown in Figure III.B.5.12. Figure III.B.5.13 shows the evolution of equity value and asset value, as well as the default point during the same period. On the vertical axis of both graphs the EDF is shown as a percentage, together with the corresponding Standard & Poor’s rating.

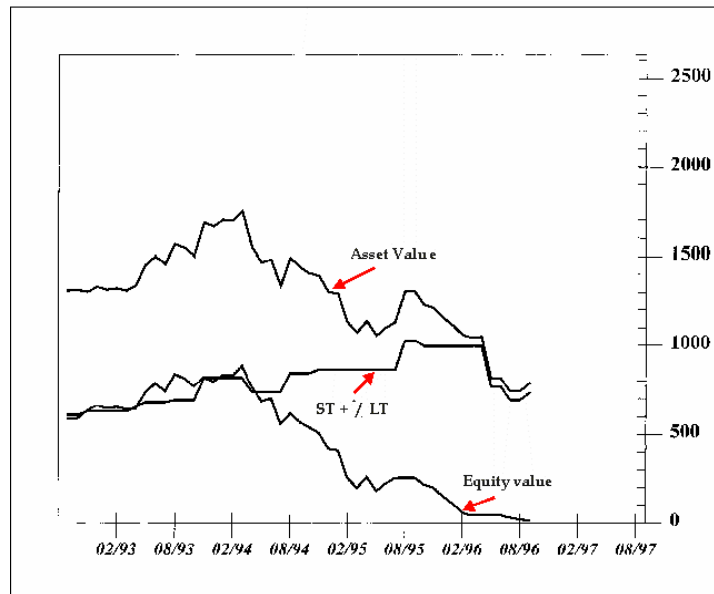
Figure III.B.5.12: EDF of a firm that defaulted versus EDFs of firms in various quartiles and the lower decile



Source: KMV Corporation

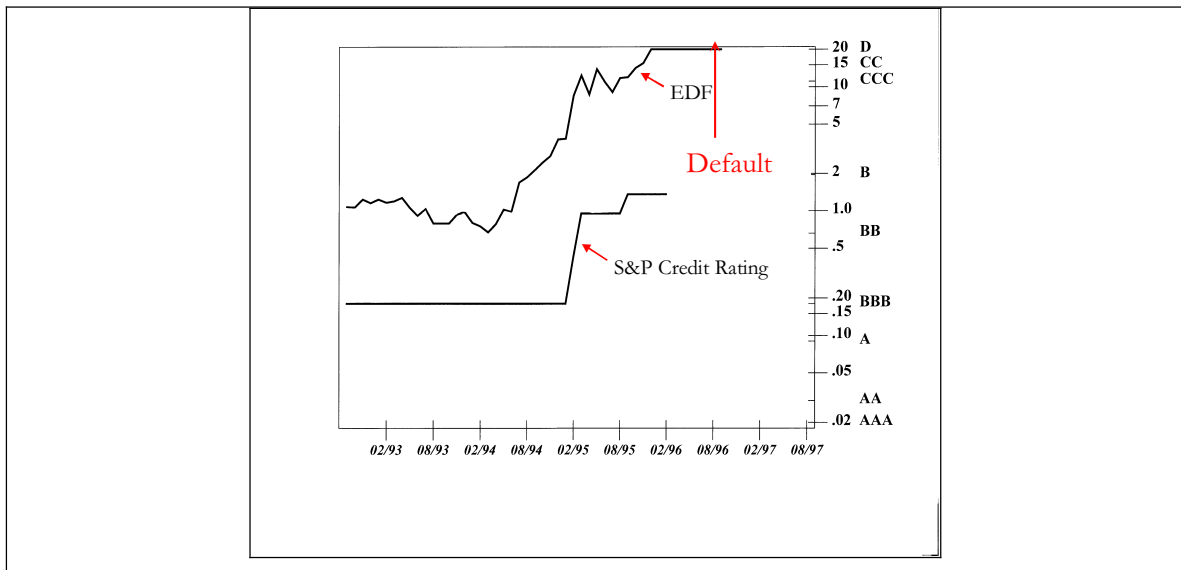
Note: The quartiles and decile represent a range of EDFs for a specific credit class.

Figure III.B.5.13: Asset value, equity value, short- and long-term debt of a firm that defaulted



Source: KMV Corporation

Figure III.B.5.14: EDF of a firm that defaulted, versus Standard & Poor’s rating



KMV has analysed more than 2000 US companies that have defaulted or entered into bankruptcy over the last 20 years. These firms belonged to a large sample of more than 100,000 company-years with data provided by Compustat. In all cases KMV has shown a sharp increase in the slope of the EDF a year or two before default. Changes in EDFs tend to anticipate – by at least one year - the downgrading of the issuer by rating agencies such as Moody’s and Standard & Poor’s (Figure III.B.5.14). Contrary to Moody’s and Standard & Poor’s historical default statistics, EDFs are not biased by periods of high or low numbers of defaults. The distance to default can be observed to shorten during periods of recession, when default rates are high, and to increase during periods of prosperity characterised by low default rates.

The loss distribution is generated in a way similar to CreditMetrics by simulating correlated defaults at the risk horizon, say one year.

III.B.5.7 The Actuarial Approach

In the structural models of default, default occurs when the asset value falls below a certain boundary such as a promised payment (e.g., the Merton, 1974, framework). By contrast, the actuarial model discussed in this section treat the firm’s bankruptcy process, including recovery, as exogenous. CreditRisk+, released in late 1997 by investment bank Credit Suisse Financial Products, is a purely actuarial model, based on mortality models of the insurance companies. This means that the probabilities of default that the model employs are based on historical statistical data of default experience by credit class.

Contrary to the structural approach to modelling default, the timing of default is assumed to take the bond-holders ‘by surprise’. Default is treated as an ‘end of game’ (stopping time) which comes as a surprise, and the probability of such a surprise is known and follow a Poisson type of distribution.

CreditRisk+ applies an actuarial science framework to the derivation of the loss distribution of a bond/loan portfolio. Only default risk is modelled; downgrade risk is ignored. Unlike the KMV approach to modelling default, there is no attempt to relate default risk to the capital structure of the firm. Also, no assumptions are made about the causes of default. It is assumed that:

1. for a loan, the probability of default in a given period, say one month, is the same as in any other month;
2. for a large number of obligors, the probability of default by any particular obligor is small, and the number of defaults that occur in any given period is independent of the number of defaults that occur in any other period.

Under these assumptions, the probability distribution for the number of defaults during a given period of time (say, one year) is represented well by a Poisson distribution:²¹

$$\Pr(n \text{ defaults}) = \frac{\mu^n e^{-\mu}}{n!} \text{ for } n = 0, 1, 2, \dots, \quad (\text{III.B.5.13})$$

where μ is the average number of defaults per year.

The annual number of defaults n is a stochastic variable with mean μ and standard deviation $\sqrt{\mu}$. The Poisson distribution has a useful property: it can be fully specified by means of a single parameter, μ . For example, if we assume $\mu = 3$ then the probability of ‘no default’ in the next year is:

$$\Pr(0 \text{ default}) = \frac{3^0 e^{-3}}{0!} = 0.05 = 5\%$$

while the probability of exactly three defaults is:

$$\Pr(3 \text{ defaults}) = \frac{3^3 e^{-3}}{3!} = 0.224 = 22.4\% .$$

²¹ In any portfolio there is, naturally, a finite number of obligors, say n ; therefore, the Poisson distribution, which specifies the probability of n defaults, for $n = 1, \dots, \infty$, is only an approximation. However, if n is large enough, then the sum of the probabilities of $n + 1, n + 2, \dots$ defaults become negligible.

However, we expect the mean default rate μ to change over time, depending on the business cycle. This suggests that the Poisson distribution can only be used to represent the default process if, as suggested by CreditRisk+, we make the additional assumption that the mean default rate is itself stochastic.

In the event of default by an obligor, the counterparty incurs a loss that is equal to the amount owed by the obligor (its exposure, i.e. the marked-to-market value, if positive – and zero, if negative – at the time of default) less a recovery amount. In CreditRisk+, the exposure for each obligor is adjusted by the anticipated recovery rate in order to calculate the ‘loss given default’. These adjusted exposures are calculated outside the model (exogenous), rather than determined by it, and are independent of market risk and downgrade risk.

In order to derive the loss distribution for a well-diversified portfolio, the losses (exposures, net of the recovery adjustments) are divided into bands. The level of exposure in each band is approximated by means of a single number.

Example III.B.5.5:

Suppose the bank holds a portfolio of loans and bonds from 500 different obligors, with exposures between \$50,000 and \$1 million.

	Notation
Obligor	A
Exposure	L_A
Probability of default	P_A
Expected loss	$\lambda_A = L_A \times P_A$

In Table III.B.5.11 we show the exposures for the first six obligors.

Table III.B.5.11: Exposure per obligor

Obligor A	Exposure (\$) (loss given default) L_A	Exposure (in \$100,000) \bar{v}_j	Round-off exposure (in \$100,000) v_j	Band j
1	150,000	1.5	2	2
2	460,000	4.6	5	5
3	435,000	4.35	4	4
4	370,000	3.7	4	4
5	190,000	1.9	2	2
6	480,000	4.8	5	5

The unit of exposure is assumed to be \$100,000. Each band $j, j = 1, \dots, m$, with $m = 10$, has an average common exposure: $v_j = \$100,000 \times j$. In CreditRisk+, each band is viewed as an independent portfolio of loans/bonds, for which we introduce the following notation:

	Notation
Common exposure in band j in units of exposure	v_j
Expected loss in band j in units of exposure	ϵ_j
Expected number of defaults in band j	μ_j

Denote by ϵ_A the expected loss for obligor A in units of exposure, i.e. $\epsilon_A = \lambda_A/L$ where in this case $L = \$100,000$. Then ϵ_j , the expected loss over a one-year period in band j , expressed in units of exposure, is simply the sum of the expected losses ϵ_A of all the obligors that belong to band j . But since by definition $\epsilon_j = v_j\mu_j$, the expected number of defaults per annum in band j may now be calculated, using $\mu_j = \epsilon_j/v_j$. Table III.B.5.12 provides an illustration of the results of these calculations.

Table III.B.5.12: Expected number of defaults per annum in each band

Band: j	Number of obligors	ϵ_j	μ_j
1	30	1.5	1.5
2	40	8	4
3	50	6	2
4	70	25.2	6.3
5	100	35	7
6	60	14.4	2.4
7	50	38.5	5.5
8	40	19.2	2.4
9	40	25.2	2.8
10	20	4	0.4

To derive the distribution of losses for the entire portfolio, we follow the three steps outlined below.

Step 1: Probability generating function for each band

Each band is viewed as a separate portfolio of exposures. The probability generating function for any band, say band j , is by definition:

$$G_j(z_j) = \sum_{n=0}^{\infty} \Pr(\text{loss} = nL) z_j^n = \sum_{n=0}^{\infty} \Pr(n \text{ defaults}) z_j^{nv_j},$$

where the losses are expressed in the unit of exposure. Since we have assumed that the number of defaults follows a Poisson distribution, we have:

$$G_j(z_j) = \sum_{n=0}^{\infty} \frac{e^{-\mu_j} \mu_j^n}{n!} z_j^{nv_j} = e^{-\mu_j + \mu_j z_j^{v_j}}.$$

Step 2: Probability generating function for the entire portfolio

Since we have assumed that each band is a portfolio of exposures that is *independent* of the other bands, the probability generating function for the entire portfolio is simply the product of the probability generating function for each band:

$$G(z) = \prod_{j=1}^m e^{-\mu_j + \mu_j z_j^{v_j}} = \exp\left(-\sum_{j=1}^m \mu_j + \sum_{j=1}^m \mu_j z_j^{v_j}\right) \quad (\text{III.B.5.14})$$

where $\mu = \sum_{j=1}^m \mu_j$ denotes the expected number of defaults for the entire portfolio.

Step 3: Loss distribution for the entire portfolio

Given the probability generating function (III.B.5.14), it is straightforward to derive the loss distribution, since

$$\Pr(\text{loss of } nL) = \frac{1}{n!} \left. \frac{d^n G(z)}{dz^n} \right|_{z=0} \text{ for } n=1,2,\dots$$

These probabilities can be expressed in closed form and depend only on ϵ_j and v_j .²²

Credit VaR is then easily derived from the above loss distribution by first computing the percentile corresponding to the confidence level, and then subtracting the expected loss from this number.

CreditRisk+ proposes several extensions of the basic one-period, one-factor model. First, the model can be easily extended to a multi-period framework. Second, the variability of default rates can be assumed to result from a number of ‘background’ factors, each representing a sector of activity. Each factor k is represented by a random variable X_k which is the number of defaults in sector k , and which is assumed to be gamma distributed. The mean default rate for each obligor is then supposed to be a linear function of the background factors, X_k . These factors are

²² See Credit Suisse (1997), p.26.

further assumed to be independent. In all cases, CreditRisk+ derives a closed-form solution for the loss distribution of a bond/loan portfolio.

CreditRisk+ has the advantage that it is relatively easy to implement. First, as we mentioned above, closed-form expressions can be derived for the probability of portfolio bond/loan losses, and this makes CreditRisk+ very attractive from a computational point of view. In addition, marginal risk contributions by obligor can be easily computed. Second, CreditRisk+ focuses on default, and therefore it requires relatively few estimates and ‘inputs’. For each instrument, only the probability of default and the exposure are required.

Its principal limitation is the same as for the CreditMetrics and KMV approaches: the methodology assumes that credit risk has no relationship with the level of market risk. In addition, CreditRisk+ ignores what might be called ‘migration risk’; the exposure for each obligor is fixed and is not sensitive to possible future changes in the credit quality of the issuer, or to the variability of future interest rates. Even in its most general form, where the probability of default depends upon several stochastic background factors, the credit exposures are taken to be constant and are not related to changes in these factors. Finally, like the CreditMetrics and KMV approaches, CreditRisk+ is not able to cope satisfactorily with non-linear products such as options and foreign currency swaps.

III.B.5.8 Summary and Conclusion

In this chapter we have presented the key features of some of the more prominent new models of credit risk measurement in a portfolio context. At first sight, these approaches appear to be very different and likely to produce considerably different loan loss exposures and VaR figures. However, analytically and empirically, these models are not as different as they may first appear. Indeed, similar arguments stressing the structural similarities have been made by several authors such as Gordy (2000) and Koyluoglu and Hickman (1999).

Comparative studies such as IIF/ISDA (2000) have stressed the sensitivity of the risk measures (expected and unexpected losses) to the key risk drivers, i.e. probabilities of default, loss given default and default correlations. This study also showed that when these models are run using consistent parameters, they produce results which fall in quite a narrow range.

References

Altman, E I, and Kishore, V (1996) Almost everything you wanted to know about recoveries on defaulted bonds. *Financial Analysts Journal*, 52(6), pp. 57–64.

- Altman, E I, and Kishore, V (1998) Defaults and returns on high yield bonds: analysis through 1997. Working Paper S-98-1, New York University Salomon Center.
- Black, F, and Scholes, M (1973) The pricing of options and corporate liabilities. *Journal of Political Economy*, 81, pp. 637–654.
- Carty, L, and Lieberman, D (1996) *Defaulted Bank Loan Recoveries*. Moody's Investors Service, Global Credit Research, Special Report.
- Credit Suisse (1997) *CreditRisk+: A Credit Risk Management Framework*. New York: Credit Suisse Financial Products.
- Crouhy, M, Galai, D, and Mark, R (2001) *Risk Management*. New York: McGraw-Hill.
- Elton, E J, and Gruber, M J (1995) *Modern Portfolio Theory and Investment Analysis*. New York: Wiley.
- Fishman G (1997) *Monte Carlo: Concepts, Algorithms, and Applications*. New York: Springer Seris in Operations Research, Springer.
- Galai, D, and Masulis, R W (1976) The option pricing model and the risk factor of stocks. *Journal of Financial Economics*, 3 (January/March), pp. 53–82
- Glasserman, P, Heidelberger, P, and Shahabuddin, P (2000) Variance reduction technique for estimating value-at-risk. *Management Science*, 46, pp. 1349–1364.
- Gordy, M (2000) A comparative anatomy of credit risk models. *Journal of Banking and Finance*, 1, pp. 119–149.
- IIF/ISDA (2000) Modeling credit risk: Joint IIF/ISDA testing program. February.
- J.P. Morgan (1997) *CreditMetrics*. Technical Document.
- Koyluoglu, H U, and Hickman, A (1998) A generalized framework for credit risk portfolio models.
- Lucas, D (1995) Default correlation and credit analysis, *Journal of Fixed Income*, 4(4), pp. 76-87. Working Paper. New York: Oliver, Wymann and Co., July.
- Merton, R C (1974) On the pricing of corporate debt: the risk structure of interest rates. *Journal of Finance*, 28, pp. 449–470.
- Rudd, A, and Clasing Jr, H K (1988) *Modern Portfolio Theory: The Principles of Investment Management*. Orienda, CA: Andrew Rudd.
- Vasicek, O (1997) Credit valuation, *Net Exposure*, 1(1).
- Wilson, T (1997a) Portfolio credit risk I, *Risk*, 10(9), pp. 111-117.
- Wilson, T (1997b) Portfolio credit risk II, *Risk*, 10(10), pp. 56-61.

III.B.6 Credit Risk Capital Calculation

Dan Rosen¹

III.B.6.1 Introduction

As discussed in Chapter III.0, the primary role of capital in a bank, apart from the transfer of ownership, is to act as a buffer to absorb large unexpected losses, protect depositors and other claim holders and provide confidence to external investors and rating agencies on the financial health of the firm. In contrast, regulatory capital refers to the minimum capital requirements which banks are required to hold, based on regulations established by the banking supervisory authorities. From the perspective of the regulator, the objectives of capital adequacy requirements are to safeguard the security of the banking system and to ensure its ongoing viability, and to create a level playing field for internationally active banks.

While the role of economic capital (EC), in general, is to act as a buffer against all the risks that may force a bank into insolvency, economic credit capital (ECC) can be viewed as a buffer against those risks specifically associated with obligor credit events such as default, credit downgrades and credit spread changes. In this chapter, we review the main concepts for estimating and allocating ECC as well as the current regulatory framework for credit capital. Chapter III.0 presents the chronology and describes the basic principles behind regulatory capital and gives an overview of the *Basel Accord*, the framework created by the Basel Committee on Banking Supervision (BCBS), which is now the basis for banking regulation around the globe. We focus in this chapter on the main principles for computing minimum regulatory credit capital requirements under the current Basel I Accord as well as under the current proposal for the new Basel accord, Basel II.

In this chapter you will learn:

- how credit portfolio models must be defined and parameterised consistently to measure ECC from a bottom-up approach;
- the basic rules for computing minimum credit capital under the Basel I Accord;
- the rules for computing minimum capital requirement for credit risk in the Basel II Accord (Pillar I) – we cover various types of exposure and comment on the special credit capital considerations under Pillar II;

¹ Algorithmics Inc.

- the Basel II treatment for internal ratings systems and probability of default estimation, as well as the minimum standards for credit monitoring processes and validation methods.

Finally, Section III.B.6.7 covers some advanced topics: the applications of risk contribution methodologies ECC allocation, as well as the shortcomings of value-at-risk (VaR) for ECC and coherent risk measures.

III.B.6.2 Economic Credit Capital Calculation

ECC acts as a buffer that provides protection against the credit risk (potential credit losses) faced by an institution. Traditionally, capital is designed to absorb *unexpected losses* up to a certain confidence level, while *credit reserves* are set aside to absorb *expected losses*. As explained in Chapter III.0, the methodologies to estimate EC at the firm level can be generally classified into *top-down* or *bottom-up* approaches. In general, top-down approaches do not readily allow the decomposition of capital into its various constituents, such as market risk, credit risk and operational risk. Bottom-up approaches provide risk-sensitive measures of capital, which allow us to understand and manage these risks better.

In this section, we discuss how credit portfolio models are applied to measure ECC from a bottom-up approach. The credit portfolio models that were described in Chapter III.B.5 provide the statistical distributions of potential credit losses of a portfolio over a given horizon. Hence, they offer the key tools to estimate the required ECC from a bottom-up approach. Chapter III.B.5 covers the fundamentals of credit portfolio models, as well as the main models used in industry.

III.B.6.2.1 Economic Capital and the Credit Portfolio Model

A credit portfolio model must be defined and parameterised consistently with the definition of economic capital used by the firm. In particular, when devising an economic capital framework, we must explicitly consider the time horizon or holding period, the definition of credit losses, and the quantile (and definition of ‘unexpected’ losses) covered by capital. The definitions of these three parameters are tightly linked to the actual definition of capital used by the firm.

III.B.6.2.1.1 Time Horizon

While trading activities tend to involve short time horizons (a few hours to a few days), credit activities generally involve longer horizons (a few months to a few years). Operational risk and insurance activities generally involve even longer horizons, spanning several years. It is common

practice for credit VaR measurement to assume a one-year horizon, although longer periods are sometimes employed.² Several reasons are cited for a one-year horizon, among them the following:

- It accords with the firm's main accounting cycle.
- It is a reasonable period over which the firm will typically be able to renew any capital depleted through losses.
- It coincides with a reasonable period over which actions can be taken to mitigate losses for various credit assets.
- Credit reviews are usually performed annually.
- The borrower might be updating its financial information only on an annual basis.

III.B.6.2.1.2 Credit Loss Definition

Accounting rules, regulatory guidance and management policy combine to determine when a loss will be recognised. Loss recognition is straightforward in some instances, for example, for trading securities that are marked-to-market regularly or operational losses that are recognised when they occur. In these instances, there is usually little question as to the function of economic capital. Management has more leeway in other instances. With regards to credit exposures, management may choose to measure economic capital against default events only (*default mode*) or against both default and credit migration events (*mark-to-market mode*):

(i) *Default-only credit losses.* In this case we are only concerned whether a loan (or other instrument) would be repaid or not. This type of measurement is currently the most prevalent, typically used for traditional (accrual accounted, hold-to-maturity) banking book activities. Credit portfolio models require several building blocks, or components, at the enterprise level. These include the following:

- Estimates of the probabilities of default (PD), loss given default (LGD), and exposures at default (EAD) for individual corporate, banking and sovereign exposures; similarly, such estimates are required for homogeneous buckets of retail or small- and medium-sized enterprise (SME) exposures.
- A portfolio model with specific assumptions about the co-dependence of
 - credit events (e.g. asset values or PD correlations)
 - exposures, LGDs and credit events.

² Note that the optimal time horizons for given portfolios might vary. However, when applied on an enterprise basis, it is important to set a common methodology to estimate the desired capital. Also, while it is most common to apply credit portfolio models in a single-step setting, the multi-step applications are becoming increasingly popular. Finally, the choices of time horizon and credit loss definition can be tightly linked.

(ii) *Mark-to-market credit losses*. In this case, the models used are closer to market risk models, capturing the losses due to defaults and the changes in mark-to-market due to deterioration in credit quality (or credit migration). Therefore, in addition to information needed for default-only models (PDs, EADs, LGDs and correlations), they require

- market pricing information (e.g. credit spreads) and
- more general information on the instruments structures for marking to market.

Mark-to-market models are more commonly used for portfolios which are not held to maturity, or for which reliable pricing information might be available, such as bond and credit derivatives portfolios.

The choice of loss definition and time horizon is tightly linked. For example, consider credit risk in lending activities. If capital is held against defaults only, it makes sense that the horizon should be the entire life of the loan. If capital is held against deterioration in value, then some time period would be selected, independent of the life of the underlying loans (e.g. one year). Both choices allow for longer-maturity instruments to carry higher capital charges.

III.B.6.2.1.3 *Quantile of the Loss Distribution*

Credit VaR should capture the ECC that shareholders should invest, in order to limit the probability of default of the firm over a given horizon. This involves defining a high, predetermined quantile of the credit loss distribution so that shareholders will be ‘highly confident’ that credit losses will not exceed this amount. The quantile, which defines the ‘confidence interval’, is chosen to provide the right level of protection to debt holders and hence achieve a desired rating, as well as providing the necessary confidence to other claim holders, such as depositors.³ The horizon and quantile are thus key policy parameters set by senior management and the board of directors. For example, a bank that wishes to be consistent with an AA rating from an external credit agency might chose a 99.97% confidence level over a one-year horizon, since the one-year default probability of an AA-rated firm is 0.03%.

III.B.6.2.2 **Expected and Unexpected Losses**

In its most common definition, economic capital is designed to absorb only *unexpected losses* (UL) up to a certain confidence level. *Credit reserves* are traditionally set aside to absorb *expected losses* (EL) during the life of a transaction. Thus, it is common practice to estimate economic capital as the chosen confidence interval minus the estimated expected EL. The key rationale for subtracting EL is that credit products are already priced such that net interest margins less non-

³ But note there can be a trade-off between achieving this objective and providing high returns on capital for shareholders.

interest expenses are at least enough to cover estimated EL (and must also cover a desired return to capital). Therefore, in this case, the credit VaR measure that is relevant to estimating EC covers only unexpected losses:

$$\text{Credit VaR}_\alpha(L) = Q_\alpha(L) - EL, \quad (\text{III.B.6.1})$$

where $Q_\alpha(L)$ denotes the $\alpha\%$ quantile (e.g. 0.03%) of the portfolio loss distribution at the horizon.

In Chapter III.0 we show that subtracting ‘EL’ as given by the expected end-of-period value from the worse-case losses represents a simplifying approximation to estimate EC. This is the approach commonly taken by practitioners and generally leads to conservative estimates.⁴ More precisely, the credit VaR measure appropriate for EC should in fact measure loss relative to the portfolio’s initial mark-to-market (MtM) value and *not* relative to the EL in its end-of-period distribution.⁵ This highlights the importance for banks of moving towards full MtM portfolio credit risk models and of establishing accurate estimates of MtM values of credit portfolios. Also, the credit VaR measure normally ignores the interest payments that must be made on the funding debt. These payments must be added explicitly to the EC measure. The estimation of the interest compensation calculation and the credit MtM value of the portfolio generally require the use of an asset pricing model.

III.B.6.2.3 Enterprise Credit Capital and Risk Aggregation

A large firm, such as a bank, acquires credit risk through various businesses and activities, including retail banking, commercial lending, bonds, derivatives and credit derivatives trading. Such an institution is likely to have one methodology for its larger commercial loans and another for its retail credits. In general, a bank may have any number of methodologies for various credit segments. If each area is modelled separately, then the amounts of ECC estimated for each area need to be combined. In making the combination, the firm needs to incorporate, either implicitly or explicitly, various correlation assumptions. The choices of aggregation can be divided into three main categories:

- Sum of stand-alone capital for each business unit or portfolio. This methodology essentially assumes perfect correlation across business lines and does not allow for diversification from them. It is consistent, in general, with one-factor portfolio models (such as the Basel II underlying portfolio model).

⁴ For a detailed discussion, see Kupiec (2002).

⁵ Capital allocation credit VaR measures in this context can be negative.

- *Ad hoc* cross-business correlation. In order to allow for some cross-business diversification, a firm might aggregate the individual stand-alone capital estimates using analytical models and simple cross-business (asset) correlation estimates.
- Full enterprise credit portfolio modelling. Current multi-factor credit models and technology allow the computation of credit loss distributions of large enterprise portfolios. This can be accomplished through semi-analytical methods (e.g. fast Fourier transforms, saddlepoint methods), or full-blown Monte Carlo simulation. Various financial institutions today are starting to apply either in-house or commercial credit portfolio models to compute ECC at the enterprise level.

III.B.6.3 Regulatory Credit Capital: Basel I

In this section you will learn the key principles of the Basel I Accord for computing minimum capital requirements for credit risk. We further discuss some of its shortcomings and the motivation for regulatory arbitrage.

III.B.6.3.1 Minimum Credit Capital Requirements under Basel I

The 1988 Basel I Accord focused mainly on credit risk, establishing minimum capital standards that linked capital requirements to the credit exposures of banks (Basel Committee of Banking Supervision 1988). Prior to its implementation in 1992, bank capital was regulated through simple, *ad hoc* capital standards. While generally prescriptive, Basel I left various choices to be made by local regulators, thus resulting in several variations of the implementation across jurisdictions. The calculation of regulatory credit capital requirements has three steps: converting exposures to credit-equivalent assets, computing loan equivalents, and applying the capital adequacy ratio.

Step 1: Credit-Equivalent Assets

The objective in this step is to express all on-balance sheet and off-balance sheet credit exposures in comparable numbers. This is achieved by converting off-balance sheet exposures into equivalent credit assets, which better reflect the amounts that could be lost if a transaction were to default. The general rules for this are as follows:

For *contingent banking book assets*, such as letters of credit, etc., asset equivalents are obtained by multiplying the exposure by a percentage, which broadly reflects the likelihood of its conversion into an actual exposure. For example, the conversion factor of an undrawn standby letter of credit is 50%.

For *derivatives in the trading book*,⁶ asset equivalents are given by the instrument's total exposure, obtained through the so-called method of add-ons:

$$\text{Total Exposure} = \text{Actual Exposure} + \text{Potential Exposure} \quad (\text{III.B.6.2})$$

$$\text{Actual Exposure} = \max(0, \text{Mark-to-Market})$$

$$\text{Potential Exposure} = \text{Notional} \times \text{Counterparty Add-on}$$

The potential exposure attempts to capture the change in value of derivatives, resulting from market fluctuations, which lead to higher credit losses should the counterparty default. Counterparty add-ons to measure potential exposures are given prescriptively, for example as in Table III.B.6.1.

Table III.B.6.1: Add-ons for derivatives exposures in Basel I

Residual maturity	Interest rate	Exchange rate and gold	Equity	Precious metals except gold	Other commodities
One year or less	0.00%	1.00%	6.00%	7.00%	10.00%
Over one year to five years	0.50%	5.00%	8.00%	7.00%	12.00%
Over five years	1.50%	7.50%	10.00%	8.00%	15.00%

Thus, for example,

- a three-year foreign exchange forward contract carries a total exposure given by its current mark-to-market plus 5% of its notional;
- an interest-rate swap with remaining nine-month maturity is deemed to carry a 0% add-on, and hence its total exposure is given by its current mark-to-market.

Basel I allows also for partial recognition of mitigation techniques such as netting when the proper agreements are in place. In this case, actual exposure is computed as the *netted* mark-to-market values of all transactions, and the total add-on is adjusted as follows:

$$\text{Netted Add-on} = \text{Gross Add-on} \times (0.4 + 0.6 \times \text{NGR Ratio}), \quad (\text{III.B.6.3})$$

where *NGR* denotes the net-to-gross ratio, that is, the netted mark-to-market of the transactions with a counterparty divided by the gross mark-to-market (the sum of all the positive mark-to-market transaction values).

⁶ This was originally defined in BCBS (1995).

Example III.B.6.1

Consider a counterparty with three derivatives transactions as shown in Table III.B.6.2. With a gross MtM of \$500m and a netted MtM of \$200m, the NGR is 0.4. Netting in this case reduces the credit equivalent for this portfolio by $(535 - 222.4)/535 = 58\%$.

Table III.B.6.2: Example – credit equivalents in the trading book

Transaction	Instrument	Residual maturity	Notional	MtM	Add-on (%)	Add-on	Credit equivalent
1	IR swap	3 years	1000	500	0.50	5	505
2	FX forward	1.5 years	500	-100	5	25	25
3	FX option	5 months	500	-200	1	5	5
	Gross			500		35	535
	Netted			200		22.4	222.4
	NGR ratio			0.4			

Table III.B.6.3: Basel I risk weights (examples)

0%	Cash Claims on central governments, central banks denominated in national currency
0, 10, 20, or 50%	Claims on domestic public-sector entities, excluding central government, and loans guaranteed by securities issued by such entities
20%	Claims on multilateral development banks Claims on banks incorporated in the OECD and loans guaranteed by OECD
50%	Loans fully secured by mortgage on residential property
100%	Claims on private sector Claims on banks incorporated outside the OECD with a residual maturity of over one year

Step 2: Risk-Weighted Assets

Risk-weighted assets (RWAs) are obtained by multiplying the exposures (or credit equivalents) by a risk weight. Risk weights broadly attempt to reflect the credit riskiness of the asset. Example risk weights are given in Table III.B.6.3. Thus, a loan to a corporate, regardless of its credit rating, carries a 100% risk weight, while a credit exposure to an OECD government has a 0% weight.

Step 3: Capital Adequacy Ratio – Minimum Capital Requirement

The minimum capital requirements are obtained by multiplying the sum of all the RWAs by the capital adequacy ratio of 8% (also referred to as the *Cook ratio*):

$$Capital = \left(\sum_k RWA_k \right) \times 8\%. \quad (\text{III.B.6.4})$$

Example III.B.6.2

Following on from Example III.B.6.1, assume that the counterparty is a UK bank. Then the risk weight is 20%, which leads to minimum capital requirements

$$\text{Min capital requirements} = \$222.4\text{m} \times 0.2 \times 0.08 = \$3.56\text{m}.$$

The reduction of regulatory capital from the application of netting for this portfolio is also 58% (as with the credit equivalent).

III.B.6.3.2 Weaknesses of the Basel I Accord for Credit Risk

As mentioned in Chapter III.0, a great strength of Basel I is the simplicity of the framework, which allowed it to be implemented in countries with different banking and accounting practices. Its simplicity also has been its major weakness, as the accord does not effectively align regulatory capital requirements closely with an institution's risk. Some criticisms with regard to credit risk include the following:

- *Lack of credit quality differentiation.* All corporate credits elicit a risk weight of 100% regardless of their credit quality. Furthermore, high-rated corporate exposures may carry much higher capital than low-rated sovereign exposures. For example, a loan to an AAA-rated corporate such as GE carries five times more capital than a loan to a bank in Korea or Turkey (20% risk weight). This indirectly provides incentives for banks to take bad credits and avoid high-rated credits (with lower returns).
- *Lack of proper maturity differentiation.* For example, revolving credit exposures with a term of less than one year do not get a regulatory capital charge. Similarly, a facility of 366 days bears the same capital charge as a long-term facility. This has led to very simple regulatory arbitrage through the systematic creation of rolling 364-day facilities.

- *Insufficient incentives for credit mitigations techniques.* The accord does not fully recognise the risk reduction achieved through credit mitigation techniques such as netting, collateral, guarantees and credit derivatives.
- *Lack of recognition of portfolio effects in credit risk.* The accord does not provide any capital benefits for diversification across assets and businesses.

III.B.6.3.3 Regulatory Arbitrage

The lack of differentiation in the accord, together with the financial engineering advances in credit risk over the last decade, have led to the widespread development of *regulatory capital arbitrage*. This refers to the process by which regulatory capital is reduced through instruments such as credit derivatives or securitisation, without an equivalent reduction of the actual risk being taken. Through regulatory arbitrage instruments, for example, banks typically transfer low-risk exposures from their banking book to their trading book, or simply place them outside the regulated banking system.

III.B.6.4 Regulatory Credit Capital: Basel II

We introduce the latest proposals of the current Basel II Accord for credit risk capital. In this section you will learn about the Pillar I rules for computing minimum capital requirement for credit capital in the Basel II Accord. We cover the standardised and internal ratings based approaches for different types of exposures and conclude with a brief comment on the special credit capital considerations under Pillar II of the new Accord.

III.B.6.4.1 Latest Proposal for Minimum Credit Capital requirements

In 1999, the BCBS issued a proposal for a new capital adequacy framework (Basel II or BIS II Accord). The third consultative paper (CP3) on the new accord was released in April 2003 (BCBS, 2003). The Final version of the accord was published in June 2004 (BCBS, 2004).⁷ The implementation of the accord will take effect between the end of 2006 and the end of 2007.

Basel II attempts to improve capital adequacy framework along two important dimensions:

- First, the development of a capital regulation that encompasses not only minimum capital requirements but also supervisory review and market discipline.
- Second, a substantial increase the risk sensitivity of the minimum capital requirements.

⁷ A small number of open issues are still to be resolved during 2004.

The reader is referred to Chapter III.0 for a general discussion on the Basel II Accord. In this section we summarise the key principles and formulae for the computation of minimum capital requirements for credit risk under Pillar I of the new accord. For greater detail, the reader is referred to the BCBS papers which can be found at www.bis.org.

As with Basel I, minimum capital requirements consist of three components:

1. definition of capital (no major changes from Basel I);
2. definition of RWA;
3. minimum ratio of capital/RWA (remains 8%).

Basel II proposes substantive changes to the treatment of RWAs for credit risk relative to Basel I. It moves away from a one-size-fits-all approach through the introduction of three distinct options for the calculation of credit risk. These approaches present increasing complexity and risk-sensitivity. Banks and supervisors can thus select the approaches that are most appropriate to the stage of development of banks' operations and of the financial market infrastructure.

Similar to Basel I, total minimum capital requirements are obtained by multiplying the risk-weighted assets by the capital adequacy ration of 8%, as in equation (III.B.6.4)⁸:

$$Capital = \left(\sum_k RWA_k \right) \times 8\% .$$

The calculation of RWAs can be done through two types of approach: the *standardised* approach and the *internal ratings based* (IRB) approach. The IRB approach has two variants, called *foundation* and *advanced* IRB.

III.B.6.4.2 The Standardised Approach in Basel II

This approach is similar to Basel I in that Basel II requires banks to slot their credit exposures into supervisory categories based on observable characteristics of the exposures (e.g. whether it is a corporate loan or a residential mortgage loan), and then establishes fixed risk weights corresponding to each supervisory category. Important differences from Basel I include the following:

- *Use of external ratings.* The standardised approach allows the use of external credit assessments to enhance risk sensitivity. The risk weights for sovereign, interbank, and

⁸ In addition, in (BCBS, 2004) the committee introduced a scaling factor. Where aggregate capital is lower than under Basel I, the new requirements must be scaled by a factor, currently estimated at 1.06, such that overall capital levels do not fall.

corporate exposures are differentiated based on external credit assessments. For sovereign exposures, these credit assessments may include those developed by OECD export credit agencies or private rating agencies. The use of external ratings for corporate exposures is optional. Where no external rating is applied to an exposure, the approach mandates that in most cases a risk weighting of 100% be used (as in Basel I). In such instances, supervisors are to ensure that the capital requirement is adequate given the default experience of the exposure type. For example, for claims on corporates the risk weights are given in Table III.B.6.4.

Table III.B.6.4: Standardised risk weights for corporate exposures

Credit assessment	AAA to AA–	A+ to A–	BBB+ to BB–	Below B–	Unrated
Risk weight	20%	50%	100%	150%	100%

- *Loans past-due.* A loan considered past-due requires a risk weight of 150%, unless a threshold amount of specific provisions has already been set aside against the loan.
- *Credit mitigants.* The approach recognises an expanded range of credit risk mitigants: collateral, guarantees, and credit derivatives. The approach expands the range of eligible collateral beyond OECD sovereign issues to include most types of financial instruments. It also sets several approaches for assessing the degree of capital reduction based on the market risk of the collateral instrument. Finally, it expands the range of recognised guarantors to include all firms that meet a threshold external credit rating.
- *Retail exposures.* The risk weights for residential mortgage exposures are reduced relative to Basel I, as are those for other retail exposures, which receive a lower risk weight than that for unrated corporate exposures. In addition, some loans to SMEs may be included within the retail treatment, subject to meeting various criteria.

Through several options, the standardised approach attempts to improve the risk sensitivity of the RWAs. Basel II also provides a ‘simplified standardised approach’, where circumstances may not warrant a broad range of options. Banks under the simplified methods must also comply with the supervisory review and market discipline requirements of Basel II.

Example III.B.6.3

Based on Table III.B.6.4, a loan to a corporate obligor, with a AA rating from an external agency, would have a capital requirement of one-fifth under the Basel II standardised approach

compared to Basel I (a risk weight of 20% versus 100%). Similarly, a loan to an A-rated obligor would see its capital requirement going to one-half (50% weight).

III.B.6.4.3 Internal Ratings Based Approaches: Introduction

In the IRB approaches, banks' internal assessments of key risk drivers serve as primary inputs to the capital calculation, leading to more risk-sensitive capital requirements. This is a substantial difference from both Basel I and the standardised approach. However, the IRB approach does not fully allow for internal portfolio models to calculate capital requirements. Instead, the risk weights (and thus the capital charges) are determined through the combination of quantitative inputs provided by banks and formulae specified by the accord. These formulae, or risk-weight functions, are based on a simple credit portfolio model, and thus align more closely with modern risk management techniques.

The IRB approach includes two variants: a *foundation* approach and an *advanced* approach. The IRB approaches cover a wide range of portfolios, with the mechanics of the calculation varying somewhat across exposure types. In the remainder of this section we

- present the key inputs and principles behind the regulatory risk-weight formulae, and
- highlight the differences between the foundation and advanced IRB approaches by portfolio, where applicable.

We present these concepts first for wholesale exposures (corporate, bank and sovereigns); then we briefly cover retail exposures and SMEs, as well as specialised lending and equity exposures.⁹

III.B.6.4.4 IRB for Corporate, Bank and Sovereign Exposures

The IRB calculation of RWAs relies on four quantitative inputs, referred to as the *risk components*:

- *Probability of default* (PD): the likelihood that the borrower will default over one year.
- *Exposure at default* (EAD): the loan amount that could be lost upon default; for loan commitments this is the amount of the facility likely to be drawn if a default occurs.
- *Loss given default* (LGD): the proportion of the exposure that will be lost if a default occurs.
- *Maturity* (M): the remaining economic maturity of the exposure.

Given these four inputs, the corporate IRB risk-weight function produces a capital requirement for each exposure. The RWA for a given exposure is given by

$$RWA = 12.5 \cdot EAD \cdot K . \quad (\text{III.B.6.5})$$

⁹ Risk weight functions represent the latest version of the Accord (BCBS, 2004)

The *capital requirement*, K , is the minimum capital per unit exposure, and is given by ¹⁰

$$K = LGD \cdot \left[N \left(\frac{N^{-1}(PD) + \sqrt{R}N^{-1}(0.999)}{\sqrt{1-R}} \right) - PD \right] \cdot MF(M, PD), \quad (\text{III.B.6.6})$$

where $N(\cdot)$ denotes the cumulative normal distribution and $N^{-1}(\cdot)$ is its inverse. The formula for K is based on a simple credit portfolio model, with some adjustments as follows:

- The term $N(\cdot)$ represents the 99.9% *default losses* of an infinitely granular homogeneous portfolio of unit exposure and 100% LGD, under a one-factor Merton-type credit model.¹¹
- The term $LGD \cdot [N(\cdot) - PD]$ denotes the *unexpected default losses*¹² of the infinitely granular portfolio (already adjusted for loss given default); that is, the 99.9% losses minus the expected losses ($EL = PD \times LGD$).
- The parameter R denotes the one-factor asset correlation for the homogeneous portfolio in the credit portfolio model. It is obtained from a calibration exercise by the BCBS:

$$R = 0.12 \left(\frac{1 - e^{-50PD}}{1 - e^{-50}} \right) + 0.24 \left(1 - \frac{1 - e^{-50PD}}{1 - e^{-50}} \right). \quad (\text{III.B.6.7})$$

R is a decreasing function of the default probability ranging from 24% to 12%, with higher-quality obligors showing higher systemic risk than lower-quality obligors. Figure III.B.6.1 gives the correlation parameter R for corporate exposures (as well as for retail).

- The final component of the capital requirement in (III.B.6.6) is the maturity function, MF , which is given by

$$MF(M, PD) = \frac{1 + (M - 2.5) \cdot b(PD)}{1 - 1.5b(PD)} \quad (\text{III.B.6.8})$$

with $b(PD) = [0.11852 - 0.05478 \cdot \log(PD)]^2$ (the function b is referred to as the *maturity adjustment*). The maturity function MF is equal to one for loans of one-year maturity. Obtained from a calibration exercise by the BCBS, MF empirically adjusts further the default losses (given by $N(\cdot)$) to the MtM losses of loans of higher maturity than one year.

¹⁰ Note that the 12.5 multiplier cancels the 8% term in the capital. This simply allows the RWA to be expressed consistently with the Basel I formulae.

¹¹ See Chapter III.B.5 for credit portfolio models and Gordy (2003) for a detailed treatment of this formula.

¹² The original formulae in CP3 (BCBS, 2003) did not subtract expected losses. After a period of consultation on the role of EL, the final version bases the risk weights exclusively on unexpected losses.

For corporate, bank and sovereign exposures the foundation IRB and advanced IRB approaches differ primarily in terms of the inputs that are provided by a bank based on its own estimates and those that have been specified by the supervisor. These differences are summarised in Table III.B.6.5.

Table III.B.6.5: Foundation and advanced IRB for corporate, bank and sovereign exposures

Data Input	Foundation IRB	Advanced IRB
Probability of default (PD)	Provided by bank based on own estimates	Provided by bank based on own estimates
Loss given default (LGD)	Supervisory values set by the Committee	Provided by bank based on own estimates
Exposure at default (EAD)	Supervisory values set by the Committee	Provided by bank based on own estimates
Maturity (M)	Supervisory values set by the Committee, or, at national discretion, provided by bank based on own estimates (with an allowance to exclude certain exposures)	Provided by bank based on own estimates (with an allowance to exclude certain exposures)

Thus, all IRB banks must provide internal estimates of PD. In addition, advanced IRB banks must provide internal estimates of LGD and EAD, while foundation IRB banks will make use of supervisory values that depend on the nature of the exposure. For example, under the foundation approach:

- LGD=45% for senior claims,
- LGD=75% for subordinated claims.

These initial LGDs are then adjusted to reflect eligible collateral and guarantees provided for each transaction. In BCBS (2004), the committee decided to adopt a more stringent definition of LGD, where LGDs must be determined based on economic downturn values, and not averages over the cycle or current values.

A major element of the IRB framework pertains to the treatment of credit risk mitigants: collateral, guarantees and credit derivatives. The LGD parameter provides a great deal of flexibility to assess the potential value of credit risk mitigation techniques. For foundation IRB

banks, therefore, the different supervisory LGD values reflect the presence of different types of collateral. Advanced IRB banks have even greater flexibility to assess the value of different types of collateral. With respect to transactions involving financial collateral, the IRB approach seeks to ensure that banks are using a recognised approach to assess the risk that such collateral could change in value, and thus a specific set of methods is provided, as in the standardised approach.

In the case of trading book exposures, Basel II outlines the same treatment of EADs as in Basel I, calculating loan equivalents through the use of add-ons and allowing partial recognition of netting and mitigation (e.g. as in equations (III.B.6.2) and (III.B.6.3)). However, practitioners and industry associations like ISDA have pointed out the limitation of such a technique in terms of its accuracy and risk sensitivity, its recognition of mitigation and natural offsets, and the over-conservative capital it demands for trading exposures, compared to loans.¹³ As industry pressure is building to allow for internal models for trading book EADs, the BCBS is currently revising this topic.¹⁴

Advanced IRB banks will generally provide their own estimates of *effective maturity* for these exposures, although there are some exceptions where supervisors can allow fixed maturity assumptions. For foundation IRB banks, supervisors can choose on a national basis whether to apply fixed maturity assumptions or to provide their own estimates of remaining maturity.

III.B.6.4.5 IRB for Retail Exposures

For retail exposures, there is only a single, advanced IRB approach and no foundation IRB alternative. Retail exposures are classified into three primary product categories, with a separate risk-weight formula for each:

- residential mortgages exposures (RMEs);
- qualifying revolving retail exposures (QRREs);
- other retail exposures (OREs).

The QRRE category refers to unsecured revolving credits, which include many credit card relationships. The other retail category refers to all other non-mortgage consumer lending, including exposures to small businesses.

¹³ For example, short of allowing for full portfolio models, Canabarro *et al.* (2003) propose to use as a loan equivalent exposure for a given counterparty the expected positive exposure (EPE) derived from an internal model (e.g. from a Monte Carlo simulation), perhaps increased by a small percentage.

¹⁴ The SEC in the USA has issued a proposed capital rule for broker-dealers, aligned with Basel II requirements, which allows for internal models; see Securities and Exchange Commission (2004).

The key inputs to the IRB retail formulae are PD, LGD and EAD, all of which are to be provided by the bank based on its internal estimates (no maturity component). In contrast to corporate exposures, these values are not estimated for individual exposures, but instead for pools of similar exposures.

Given these three inputs, the retail IRB risk-weight function produces a specific capital requirement for each homogeneous pool of exposures. The risk-weighted assets for a given exposure are also given by expression (III.B.6.5).

The formula for the capital requirement K for all three retail product categories is

$$K = LGD \cdot \left[N \left(\frac{N^{-1}(PD) + \sqrt{R} N^{-1}(0.999)}{\sqrt{1-R}} \right) - PD \right] \quad (\text{III.B.6.9})$$

where again N denotes the cumulative normal distribution and N^{-1} is its inverse. The term $N(\cdot) - PD$ is the same as for corporate exposures: the 99.9% *unexpected* default losses of an infinitely granular homogeneous portfolio of unit exposure and 100% LGD. As there is no maturity adjustment for retail exposures, capital only covers default risk.

The parameter R denotes the one-factor asset correlation for the homogeneous portfolio and is different for each product category, according to a calibration exercise by the BCBS:

- residential mortgages $R = 15\%$
- *QRREs*¹⁵ $R = 4\%$
- other retail $R = 0.03 \left(\frac{1 - e^{-35PD}}{1 - e^{-35}} \right) + 0.16 \left[1 - \frac{1 - e^{-35PD}}{1 - e^{-35}} \right]$

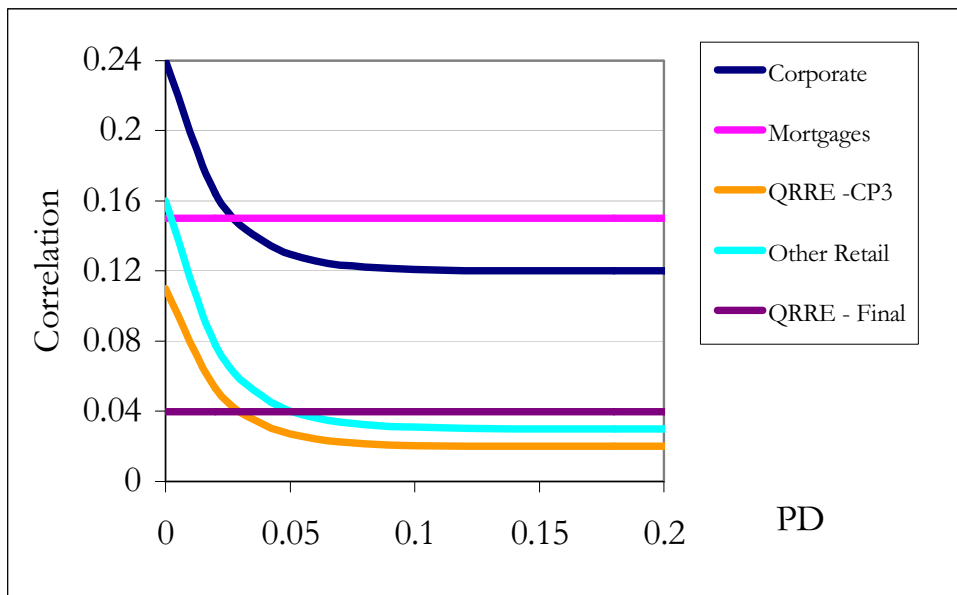
Figure III.B.6.1 gives the correlation parameter R for retail as well as corporate exposures. Correlations are in practice smaller for retail than wholesale. Corporate exposures have declining correlations ranging from 24% to 12% . Retail correlations are generally smaller, flat at 4% and 15% for QRREs and mortgages, and from 16% to 3% for other retail.

¹⁵ In CP3, the correlation for QRREs was originally given by

$$R = 0.02 \left(\frac{1 - e^{-50PD}}{1 - e^{-50}} \right) + 0.11 \left[1 - \frac{1 - e^{-50PD}}{1 - e^{-50}} \right]$$

Industry groups recommended revisions for the retail correlation curves based on best practices (e.g. RMA, 2003), suggesting that correlations for retail exposures were lower and not dependent on PD, as with wholesale exposures. In response, the final version of the accord uses a flat 4% correlation.

Figure III.B.6.1: Correlation parameter, R



In (III.B.5.9), as with wholesale, $EL = LGD \times PD$ is excluded from minimum capital, but is to be monitored through rules that determine the value of a bank's total regulatory capital.¹⁶

III.B.6.4.6 IRB for SME Exposures

Exposures not classified as purely 'retail', but with turnover of less than €50m, are classified as SMEs. They are further divided into:

- retail SMEs, where exposures are less than €1m (they may be treated as retail exposures if they are managed by the bank in the same way as retail exposures);
- corporate SMEs, where the exposure is more than €1m.

SMEs that fall into the corporate approach will apply the corporate IRB risk-weight formula, with an optional firm-size adjustment, which leads to a discount in minimum capital. The value of the adjustment is a function of turnover of the borrower (up to €50 million). The average discount will be about 10%, and can range between 1% and 20%. Within this range, a mid-sized firm with a 2% PD would be weighted at 100%.

Retail SMEs can be treated using the retail IRB formulae. This requires showing that their exposures are below the €1m threshold and that the bank indeed manages them as aggregated retail (lower-risk, small, diversified loans managed on a pooled basis).

¹⁶ In this sense, banks will be required to compare loan loss provisions (general and specific) to EL, and deduct any shortfall from its capital; excess provisions are credited through Tier 2 capital.

III.B.6.4.7 IRB for Specialised Lending and Equity Exposures

Specialised lending is associated with the financing of individual projects where the repayment is highly dependent on the performance of the underlying pool or collateral. Two options are given to treat these exposures:

- For all but one of the specialised lending subcategories, if banks can meet the minimum criteria for the estimation of the relevant inputs, they can use the corporate IRB framework to calculate risk weights. For ‘high-volatility commercial real estate’ (HVCRE), IRB banks that can estimate the required inputs will use a separate risk weight that is more conservative than the general corporate risk weight.
- Since the hurdles for meeting these criteria for this set of exposures may be more difficult in practice, CP3 also includes an additional option that only requires that a bank classify such exposures into five distinct quality grades, and provide a specific risk weight for each of these grades.

IRB banks must separately treat their equity exposures. Two distinct approaches are given:

- The first approach builds on the PD/LGD approach for corporate exposures and requires banks to provide their own PD estimates for the associated equity exposures. This approach, however, mandates the use of a 90% LGD value and also imposes various other limitations, including a minimum risk weight of 100% in many circumstances.
- The second approach provides the opportunity to model the potential decrease in the market value of equity holdings over a quarterly holding period. A simplified version of this approach with fixed risk weights for public and private equities is also included.

III.B.6.4.8 Comments on Pillar II

Pillar II of the Basel Accord (supervisory review) is based on a series of guiding principles which point to the need for banks to assess their capital adequacy positions relative to their overall risks, and for supervisors to review and take appropriate actions in response to those assessments. Important new components of Pillar II for credit risk include the treatment of the following:

- *Stress testing.* Banks adopting the IRB approach to credit risk will be required to perform a meaningfully conservative stress tests of their own design to estimate the potential increases in capital requirements during a stress scenario. Stress-test results are to be used as a means of ensuring that banks hold a sufficient capital buffer to protect against adverse or uncertain economic conditions. To the extent that there is a capital shortfall, supervisors may, for example, require a bank to reduce its risks or increase its existing capital resources to cover its minimum capital requirements plus the results of a recalculated stress test.

- *Concentration risk.* Minimum capital risk weights assume that the portfolio is large and well diversified. Within Pillar II, banks will need to assess the degree of concentration risk they incur.
- *Residual risks arising from the use of collateral, guarantees and credit derivatives,* as well as specific *securitisation* exposures (such as significant risk transfer and considerations related to the use of call provisions and early amortisation features).

III.B.6.5 Basel II: Credit Model Estimation and Validation

In this section we broadly introduce the Basel II methodology for probability of default estimation, the distinction between point-in-time and through-the-cycle ratings, the minimum standards for credit monitoring processes and the validation methods.

III.B.6.5.1 Methodology for PD Estimation

Basel II requires that PDs be derived from a two-stage process:

1. Each obligor must be classified into a *risk bucket*, corresponding to its internal rating grade. Obligor within a bucket share the same credit quality as assessed by the bank's internal credit rating system, which must be based on clear rating criteria.
2. A *pooled PD* is calculated for each bucket. For minimum capital calculations, this PD is assigned to each obligor in a given bucket. Pooled PDs must be long-run averages of one-year realised default rates for borrowers in the bucket.

III.B.6.5.2 Point-in-Time and Through-the-Cycle Ratings

Since pooled PDs are assigned to rating buckets (rather than to individual obligors directly), they can differ meaningfully from individual obligors' PDs that result from a forecasting model. Thus, the credit ratings approach can have a substantial effect on the on the minimum capital requirements. In this sense, credit rating approaches can be classified into two categories (see, for example, BCBS, 2000):

- In a *point-in-time* (PIT) rating approach, obligors are classified into rating grades based on the best available current credit quality information. The internal rating reflects an assessment of the borrower's current condition and/or most likely future condition over the time horizon. Thus, a rating changes as the borrower's conditions changes over the credit/business cycle. In general, PIT ratings tend to rise during business expansions as obligors' creditworthiness improve and tend to fall during recessions.
- In a *through-the-cycle* (TTC) rating approach, obligors are assigned to rating grades based on their ability to remain solvent over a full business cycle or during stress events. A borrower's riskiness assessment is thus based on a worst-case, 'bottom of the cycle'

scenario. Since they emphasise stress conditions, borrowers' ratings would tend to be more stable over the credit/business cycle.

In practice, there is large variation between banks' rating approaches. Furthermore, the terms PIT and TTC are often defined poorly and used differently across institutions. While not explicitly requiring that rating systems be PIT or TTC, Basel II does hint at a preference for TTC approaches.¹⁷

III.B.6.5.3 Minimum Standards for Quantification and Credit Monitoring Processes

Basel II gives banks great flexibility in determining how obligors are assigned to buckets, but it establishes minimum standards for their credit monitoring processes. Banks can rely on their own internal data, or data derived from external sources as long as they can demonstrate the relevance of such data to their own exposures. In practical terms, banks will be expected to have in place a process that enables them to collect, store and utilise loss statistics over time in a reliable manner.

Other minimum standards include the following:

- Internal rating systems should accurately and consistently differentiate degrees of risk.
- Banks must define clearly and objectively the criteria for their rating categories.
- A strong control environment must be in place to ensure that banks' rating systems perform as intended and that the resulting ratings are accurate.
- Banks must have independent and transparent ratings processes and internal reviews.

III.B.6.5.4 Validation of Estimates

Basel II requires that banks must have a robust system in place to validate the accuracy and consistency of rating systems, processes, and the estimation of all relevant risk components. While the accord does not go into methodological details, regulators broadly describe two empirical approaches for validating PDs:

- *Benchmarking* involves comparing reported PDs for similar obligors across banks and other external systems. Thus, banks that do not estimate PDs effectively or systematically misrepresent them will report substantial differences with respect to their peers.

¹⁷ From BCBS (2004, paragraph 415): 'A borrower rating must represent the bank's assessment of the borrower's ability and willingness to contractually perform despite adverse economic conditions or the occurrence of unexpected events. For example, a bank may base rating assignments on specific, appropriate stress scenarios. Alternatively, a bank may take into account borrower characteristics that are reflective of the borrower's vulnerability to adverse economic conditions or unexpected events, without explicitly specifying a stress scenario. The range of economic conditions that are considered when making assessments must be consistent with current conditions and those that are likely to occur over a business cycle within the respective industry/geographic region.'

- *Backtesting* compares the pooled PD for a grade with the actual observed (out of sample) default frequencies for that grade. If a grade's pooled PD is truly an estimate of the long-run average of the grade's observed yearly default frequencies, over time the two should converge. However, in practice, this convergence could take many years, thus providing important technical challenges.

III.B.6.6 Basel II: Securitisation

A *securitisation* is a financial structure where cash flows from an underlying pool of exposures are used to service one or more stratified positions, or *tranches*, reflecting different degrees of credit risk. Payments of the structure depend on the performance of the underlying exposures. In a *traditional securitisation*, the underlying pool commonly contains standard credit instruments such as bonds or loans. In contrast, a *synthetic securitisation* uses credit derivatives or guarantees in a funded (e.g. credit-linked notes) or an unfunded (e.g. credit default swaps) way. Common examples of securitisation structures are collateralised bond obligations, collateralised loan obligations and asset-backed securities. Securitisation by its very nature relates to the transfer of risks associated with the credit exposures of a bank to other parties. In this respect, it provides better risk diversification and contributes to enhancing financial stability. See Chapter I.B.6, and Section I.B.6.6 in particular, for more details on these securitisation structures.

Some securitisations have enabled banks under the current Basel I Accord to avoid maintaining capital commensurate with the risks to which they are exposed. This is commonly referred to as *regulatory arbitrage*: the avoidance of minimum regulatory capital charges through the sale or securitisation of a bank's assets for which the true risk (and hence economic capital) is much lower than regulatory capital. In contrast, Basel II provides a specific treatment for securitisation, which requires banks to look to the economic substance of a securitisation transaction when determining the appropriate capital requirement in both the standardised and IRB treatments.

Under the standardised approach, banks must assign risk weights prescribed by the accord according to various criteria such as a facility type and its external rating (if available). Banks that apply the standardised approach to the type of exposures securitised must also use it under the securitisation framework. Some examples of capital charges under the standardised approach are as follows:

- A structure with a long-term rating of BBB is assigned a capital of 8% times its notional while an A-rated is assigned 4% (100% and 50% risk weights, respectively).
- Most unrated structures have a capital factor of 100% (exceptions might be some most senior tranches).

- Eligible liquidity facilities satisfying some basic criteria might be assigned a lower capital by means of a credit conversion factor (CCF) which is less than 100% (total capital is the product of the notional, the capital factor and the CCF).

Banks applying the IRB approach for the type of exposures securitised must also apply it to securitisations. The IRB approach proposes three methods for calculation:

- *Ratings-based approach* (RBA). In this case capital factors are tabulated based on a tranche's credit rating and thickness, as well as the granularity of the underlying pool. Thus, for example, a AAA thick tranche with granular pool is assigned 0.56% capital (a 7% risk weight), while a BB+ tranche gets 20% capital (250% risk weight).
- *Internal assessment approach* (IAA). Subject to meeting various operational requirements, a bank may use its internal assessments of the credit quality of the securitisation exposures that it extends to asset-backed commercial paper programmes (e.g. liquidity facilities and credit enhancements). The internal assessment is then mapped to an equivalent external rating, which is used to determine the appropriate risk weights under the RBA.
- *Supervisory formula approach* (SFA). This approach is an attempt to provide a closed-form capital charge, based on a bottom-up risk assessment of the structure. The methodology used to determine the formula is based on similar mathematical modelling of the problem to that used to derive the IRB capital charge of individual exposures (see equation (III.B.6.6)). In essence, the SFA is based on five bank-supplied parameters:
 - K_{irb} – the capital charge of the underlying pool of exposures, had the assets not been securitised;
 - L – the credit enhancement supporting a given tranche (i.e. how big is the buffer absorbing credit losses, before they hit the tranche);
 - T – the thickness of the tranche;
 - N – the effective number of exposures in the pool;
 - LGD – the exposure-weighted average LGD of the pool.

The reader is further referred to various BCBS documents on securitisation; those interested in the mathematics of the SFA are referred to Gordy and Jones (2003) and Pykhtin and Dev (2002 and 2003).

Basel II prescribes the use of these three approaches in a hierarchical manner. The RBA must be applied to securitisation exposures that are rated, or where a rating can be. Where an external or an inferred rating is not available, either the SFA or the IAA must be applied. Securitisation exposures to which none of these approaches can be applied must be deducted.

III.B.6.7 Advanced Topics on Economic Credit Capital¹⁸

In this section we review

- the application of credit risk contribution methodologies for ECC allocation, and
- the shortcomings of VaR for ECC and coherent risk measures.

III.B.6.7.1 Credit Capital Allocation and Marginal Credit Risk Contributions

In this section, we briefly highlight key issues on the application of the methodologies introduced in Section III.0.4.2 for computing marginal credit risk contributions for ECC allocation.

In addition to computing the total ECC for portfolio, it is important to develop methodologies to attribute this capital *a posteriori* to various sub-portfolios such as the firm's activities, business units and even individual transactions and allocate it *a priori* in an optimal fashion, to maximise risk-adjusted returns. EC allocation down the portfolio is required for management decision support and business planning, performance measurement and risk-based compensation, pricing, profitability assessment and limits, building optimal risk–return portfolios and strategies.

There is no unique method to allocate ECC. Section III.0.4.2 classifies EC contributions into *stand-alone* contributions, *incremental* contributions, and *marginal* contributions.¹⁹ Every methodology has its advantages and disadvantages, and might be more appropriate for a particular managerial application.

The most common approach used today to attribute ECC on a diversified basis is based on the marginal contribution to the volatility (or standard deviation) of the portfolio losses. Such allocations are generally ineffective for credit risk, since loss distributions are far from normally distributed, producing inconsistent capital charges, and in some cases a loan's capital charge can even exceed its exposure (see Praschnik *et al.*, 2001; Kalkbrener *et al.*, 2004).

Given the definition of ECC, the natural choice for allocating capital is the risk contributions to a VaR-based measure. However, VaR has several shortcomings since it is not a coherent risk measure (see Section III.B.6.7.2). Specifically, while VaR is sub-additive for normal distributions, this is not true in general. This limitation is particularly relevant for credit losses, which may be far from normal and not even smooth.²⁰ Furthermore, the pointwise nature of VaR has generally lead to difficulties in computing accurate and stable risk contributions with simulation. Recently,

¹⁸ This section is not mandatory for the exam, but is added for completeness.

¹⁹ The reader is reminded that there is currently no universal terminology for these methodologies in the literature.

²⁰ The discreteness of individual credit losses leads to non-smooth profiles and marginal contributions.

several authors have proposed the use of expected shortfall (ES) for attributing ECC (see, for example, Kalkbrener *et al.*, 2004).

As a coherent risk measure, ES represents a good alternative both for measuring and allocating capital. In particular, ES yields additive and diversifying capital allocations. This requires, however, a modification of the standard interpretation of ECC to act as ‘a buffer for an expected loss conditional on exceeding a certain quantile’.

As explained in Section III.0.4, marginal contributions require the computation of a derivative of the risk measure (see equation (III.0.16)). The general theory behind the definition and computation of these derivatives in terms of quantile measures (e.g. VaR, ES) has recently been developed (see Gouriéroux *et al.*, 2000; Tasche, 2000, 2002). Several semi-analytical approaches have also recently been proposed for VaR or ES contributions (e.g. Martin *et al.*, 2001; Kurth and Tasche, 2003). While computing the conditional expectations can be challenging when credit losses are estimated from a Monte Carlo simulation, various methodologies have been devised in recent years for VaR and ES contributions in simulation models (see Kalkbrener *et al.*, 2004; Hallerbach, 2003; Mausser and Rosen, 2004). Simulation provides the flexibility to support more realistic credit models, which include diversification through multiple factors, more flexible co-dependence structures, multiple asset classes and default models, stochastic (correlated) modelling of exposures, and loss given default.

III.B.6.7.2 Shortcomings of VaR for ECC and Coherent Risk Measures

In its common interpretation, ECC is a buffer that provides protection against potential credit losses, at a confidence level that is less than 100% (e.g. 99.9%). The confidence level is consistent with the desired credit rating of the firm. This leads to measures of capital that reflect a given quantile of a credit loss distribution.

Given this definition, VaR is an intuitive measure for ECC. However, VaR has several shortcomings since it is not a coherent risk measure (in the sense of Artzner *et al.*, 1999). In particular, VaR is not sub-additive in general.

A risk measure ϱ is said to be sub-additive if

$$\varrho(X + Y) \leq \varrho(X) + \varrho(Y) \tag{III.B.6.10}$$

for any two portfolios X and Y . Thus, sub-additivity is a property of risk measures required to account for portfolio diversification.

While VaR is always sub-additive for normal loss distributions, in more general cases the total portfolio VaR might be higher than the sum of stand-alone VaRs. This is particularly relevant to credit loss distributions, which are far from normal and not smooth, given the discreteness of individual credit losses

Example III.B.6.4

Consider a simple one-year BBB loan with a notional of \$100. The obligor has a PD of 0.5% and a 50% LGD. Expected losses are $\$100 \times 0.5 \times 0.005 = \0.25 . However, the 99% VaR is 0 (we are more than 99% certain that we will not incur a loss). This leads to a 'negative' unexpected loss and, thus, the stand-alone capital of this position is negative.

Now consider a portfolio that invests \$100 equally in 10 one-year loans to different BBB obligors. Assume obligor defaults are independent and a 50% LGD for all of them. In this case, EL is still \$0.25. There is now also a 95.11% probability of no defaults and a 99.89% chance that there is at least one defaulted loan. Thus, the 99% VaR is equal to \$5 ($\$10 \text{ notional} \times 50\% \text{ LGD}$). Based on VaR, the total credit capital to support this portfolio is \$4.75. However the stand-alone VaR of each loan is zero (thus the stand-alone capital of each loan remains, in principle, $-\$0.25$).

The theory of coherent risk measures (Artzner *et al.*, 1999) is well developed and has become popular among academics and practitioners. Coherent risk measures such as expected shortfall present a good alternative both for measuring and allocating capital. This requires, of course, the modification of the standard definition and interpretation of capital as a buffer to cover $\alpha\%$ losses, to one where the buffer would cover the 'expected losses conditional on reaching an $\alpha\%$ loss'.

III.B.6.8 Summary and Conclusions

This chapter reviews the main concepts for estimating and allocating ECC as well as the current regulatory framework for credit capital. Credit portfolio methodologies are the key tools to compute ECC from a bottom-up approach. Today, they are used broadly by practitioners for estimating ECC and managing credit risk at the portfolio level. Furthermore, various institutions are starting to apply credit portfolio frameworks to compute and manage ECC at the enterprise level.

Credit portfolio models must be defined and parameterised consistently with the ECC definition of the firm. This definition includes the time horizon, the type of credit loss (default only or mark-to-market) and the confidence level (or quantile) of the loss distribution. Since ECC is

designed to absorb unexpected losses up to a certain confidence level, it is commonly estimated by a VaR-type measure (at the defined confidence level) which subtracts expected losses. We further address some potential shortcomings of VaR for measuring risk as well as for allocating capital, and discuss other measures such as expected shortfall.

In the past, regulatory credit capital has differed significantly from ECC. However, the new Basel II Accord for banking regulation has introduced a closer alignment of regulatory credit capital with current best-practice credit risk management and ECC measurement. While today falling short of allowing the use of credit portfolio models to estimate regulatory credit capital, Basel II has introduced various approaches for minimum capital requirements of increased complexity and alignment with the credit riskiness of an institution. In particular, for the first time, it allows banks to use internal models for estimating key credit risk components (PDs, exposures and LGDs). Furthermore, the IRB risk-weight formulae are based on solid credit portfolio modelling principles. Finally, with its three-pillar foundation, Basel II focuses not only on the computation of regulatory capital, but also on a holistic approach to managing risk at the enterprise level.

References

Artzner, P, Delbaen, F, Eber, J-M, and Heath, D (1999) Coherent measures of risk, *Mathematical Finance*, 9(3), pp. 203–228.

Basel Committee on Banking Supervision (1988) International convergence of capital measurement and capital standards. Available at <http://www.bis.org>

Basel Committee on Banking Supervision (1995) Basel capital accord: treatment of potential exposure for off-balance-sheet items. Available at <http://www.bis.org>

Basel Committee on Banking Supervision (2000) Range of practice in banks' internal ratings systems. Discussion paper, available at <http://www.bis.org>

Basel Committee on Banking Supervision (2003) The new Basel capital accord: Consultative document. Available at <http://www.bis.org>

Basel Committee on Banking Supervision (2004) International convergence of capital measurement and capital standards: A revised framework. Available at <http://www.bis.org>

Canabarro, E, Picoult, E, and Wilde, T (2003) Analysing counterparty risk. *Risk*, September, pp. 117–122.

Gordy, M (2003) A risk-factor model foundation for ratings-based bank capital rules, *Journal of Financial Intermediation*, 12(3), pp. 199–232.

Gordy, M, and Jones, D (2003) Random tranches, *Risk*, March, pp. 78–83.

Gouriéroux, C, Laurent, J-P, and Scaillet, O (2000) Sensitivity analysis of values at risk, *Journal of Empirical Finance*, 7(3–4), pp. 225–245.

- Hallerbach, W G (2003) Decomposing portfolio value-at-risk: a general analysis, *Journal of Risk*, 5(2), pp. 1–18.
- Kalkbrener, M, Lotter, H, and Overbeck, L (2004) Sensible and efficient capital allocation for credit portfolios, *Risk*, January, pp. S19–S24.
- Kurth, A, and Tasche, D (2003) Contributions to credit risk. *Risk*, March, pp. 84–88.
- Kupiec, P (2002) Calibrating your intuition: Capital allocation for market and credit risk. IMF Working Paper WP/02/99, available at <http://www.imf.org>
- Martin, R, Thompson, K, and Browne, C (2001) VAR: who contributes and how much? *Risk*, August, pp 99–102.
- Mausser, H, and Rosen, D (2004) Scenario-based risk management tools. In S W Wallace and W T Ziemba (eds), *Applications of Stochastic Programming*. Philadelphia: SIAM.
- Praschnik, J, Hayt, G, and Principato, A (2001) Calculating the contribution, *Risk*, 14(10), pp. S25–S27.
- Pykhtin, M, and Dev, A (2002) Credit risk in asset securitisations: an analytical model, *Risk*, May, pp. S16–20.
- Pykhtin, M, and Dev, A (2003) Coarse-grained CDOs. *Risk*, January, pp. 113-116.
- Risk Management Association (2003) Retail credit economic capital estimation – best practices. Working Paper, Risk Management Association, available at <http://www.rmahq.org>
- Securities and Exchange Commission (2004) Proposed rule: Alternative net capital requirements for broker-dealers that are part of consolidated supervised entities. 17 CFR Part 240. <http://www.sec.gov/rules/proposed/34-48690.htm>
- Tasche, D (2000) Conditional expectation as quantile derivative. Working paper, Technische Universität München.
- Tasche, D (2002) Expected shortfall and beyond. Working paper, Technische Universität München.

III.C.1 The Operational Risk Management Framework

Michael K. Ong¹

In this chapter I provide a brief outline of how to establish an operational risk management framework within an institution. Operational risk has received a lot of attention recently although it is not an entirely new field of risk management. Many of the biggest losses in the financial industry and the corporate arena can be attributed, in one way or another, to operational risk failures. The chapter begins by highlighting some of the better-known losses in the recent past and argues why it is important for individual institutions to define what operational risk means to them. The Basel II proposals, scheduled for promulgation in 2006, have also provided some guidance (primarily to banking institutions) on the types of operational risk failure and their associated loss event types. The chapter then discusses the goals and scope of an operational risk management framework. It outlines the key components of operational risk and presents some useful tools, e.g., the risk catalogue and risk scorecard, for identifying specific operational risk failures. Finally, it explains how to make the risk assessment process work through the involvement of senior management and every business unit within the institution.

III.C.1.1 Introduction

Operational risk management has become increasingly important for financial institutions over the past several years. The need for a better understanding of operational risk is driven primarily by two factors, namely, the growing sophistication of financial technology and the rapid deregulation and globalisation of the financial industry. These factors contribute to the increasing complexity of banking activities and, therefore, heighten the operational risk profile of the financial services industry. Over the past few years, a significant number of high-impact and high-profile losses, some leading to the demise of once revered, well-respected institutions, have pointed consistently to failure in operational risk management.

This seemingly sudden awareness of operational risk management is quite ironic considering that operational risk has always been an integral risk associated with doing business. ‘Operational risk is as old as the banking industry itself’, the rating agency Fitch reports, ‘and yet, the industry has only recently arrived at a definition of what it is’. The report goes on to say that ‘in its rating analysis of banks, Fitch will be looking for evidence of a clearly articulated definition of operational risk, examining the quality of an organization’s structure and operational risk culture,

¹ Professor of Finance and Executive Director of the Center for Financial Markets, Stuart Graduate School of Business, Illinois Institute of Technology. The author wishes to extend his sincerest thanks to the editors, Carol Alexander and Elizabeth Sheedy, for their careful editing of this chapter.

the development of its approach to the identification and assessment of key risks, data collection efforts, and overall approach to operational risk quantification and management' (Ramadurai *et al.*, 2004). In addition, Moody's believes that 'operational risk management improves the quality and stability of earnings, thereby enhancing the competitive position of the bank and facilitating its long-term survival' (Moody's Investor's Service, 2003). Moody's goes on to comment that: 'The control of operational risk is fundamentally concerned with good management, which involves a tenacious process of vigilance and continuous improvement. This is a value-adding activity that impacts, either directly or indirectly, on bottom-line performance. It must, therefore, be a key consideration for any business. Since operational risk will affect credit ratings, share prices, and organisational reputation, analysts will increasingly include it in their assessment of the management, their strategy and the expected long-term performance of the business.' Thus rating agencies are now clearly interested in how financial institutions manage their operational risk. In fact, how institutions manage their operational risks is likely to influence how they will be rated by the rating agencies.

Against the background of greater complexity and opaqueness in the banking industry due to technological advancement, the Basel Committee for Banking Supervision (2003) cites the emergence of new forms of risk that require attention immediately:

'Developing banking practices suggest that risks other than credit, interest rate and market risk can be substantial. Examples of these new and growing risks faced by banks include:

- If not properly controlled, the greater use of more highly automated technology has the potential to transform risks from manual processing errors to system failure risks, as greater reliance is placed on globally integrated systems;
- Growth of e-commerce brings with it potential risks (e.g., internal and external fraud and system security issues) that are not yet fully understood;
- Large-scale acquisitions, mergers, de-mergers and consolidations test the viability of new or newly integrated systems;
- The emergence of banks acting as large-volume service providers creates the need for continual maintenance of high-grade internal controls and back-up systems;
- Banks may engage in risk mitigation techniques (e.g., collateral, credit derivatives, netting arrangements and asset securitisations) to optimise their exposure to market risk and credit risk, but which in turn may produce other forms of risk (e.g. legal risk); and

- Growing use of outsourcing arrangements and the participation in clearing and settlement systems can mitigate some risks but can also present significant other risks to banks.

The diverse set of risks listed above can be grouped under the heading of “operations risk”.

The emergence of the types of risks listed above by the Basel Committee forms the basis for regulatory pressure currently felt by many major financial institutions. The Financial Services Authority (2003) reported that, even as late as in mid-2003, the financial industry was still in the early stages of developing operational risk frameworks. In its survey, the FSA reported that ‘a majority of firms stated that their primary motivation for developing the operational framework was increased regulatory focus, with regulation a more significant driver in smaller firms than in major financial groups’.

Should the impetus for developing a sound operational risk management framework be driven primarily by emerging concerns raised by rating agencies or the threat of greater scrutiny by regulatory authorities? In fact, should operational risk attain the limelight it is currently basking under because of the impending capital charge being deliberated in Basel II? I think not. These motivations in isolation are unlikely to lead to successful implementation of an operational risk management function. A coherent, sound and successful operational risk management framework can only come from an internal realisation and desire amongst senior management that this is a value-adding activity that ultimately impacts the bottom line of the institution.

III.C.1.2 Evidence of Operational Failures

Table III.C.1.1 lists some of the largest derivatives losses on the Street during the 1990s. In all cases the losses are attributable, at least in part, to operational risk. Losses resulted from flaws in the risk management framework of the institutions concerned.

One of the most dramatic and well-documented derivatives losses was the collapse in 1995 of Barings, Britain’s oldest merchant bank (200 years!). One person (Nick Leeson), based in Singapore (several thousand miles away from corporate headquarters), managed to circumvent internal systems over an extended period of time to hatch and hide his trading schemes. It ultimately resulted in over \$1.2bn of losses. This all pointed to senior management’s abject failure to institute proper managerial, financial and operational control over the institution. Since the bank’s risk management and control functions were very weak, the system of checks and balances failed at several operational and managerial junctures and in more than one location where the bank operated. The Barings debacle is not the story of just one single solitary rogue trader, but

rather the breakdown of an entire organisation that had failed to exercise sufficient oversight and control of its *people* at all levels, its lack of clear directions and accountability for the *processes* within the bank, and the failure of *technology* to detect trading and booking anomalies for an extended period of time.

Yet all of these derivatives losses combined pale in comparison to the S&L crisis (\$150bn) of the 1980s, the ‘non-performing’ real estate loans of Japanese banks (\$500bn) in the early 1990s, the Credit Lyonnais bankruptcy (\$24bn) due to bad debt in 1996, and the more recent asset management frauds at Deutsche Morgan Grenfell, Jardine Fleming, etc.²

The early 2000s witnessed the multi-billion dollar collapse of Enron, WorldCom, Tyco, Parmalat, and many other fallen angels which, in more ways than one, brought about a sense of urgency for better corporate governance. Corporate governance in essence calls for greater accountability of senior management in an effort to combat corporate fraud. And corporate fraud is one common instance of operational risk failure.

Table III.C.1.1: Publicly disclosed derivatives losses in the 1990s

Company/Entity	Loss Amount (\$m)	Area of loss
Air Products	113	<i>Leverage & currency swaps</i>
Askin Securities	600	<i>Mortgage-backed securities</i>
Baring Brothers	1240.5	<i>Options</i>
Cargill (Minnetonka Fund)	100	<i>Mortgage derivatives</i>
Codelco Chile	200	<i>Copper & precious metals futures and forwards</i>
Glaxo Holdings PLC	150	<i>Mortgage derivatives</i>
Long Term Capital Management	4000	<i>Currency & interest rate derivatives</i>
Metallgesellschaft	1340	<i>Energy derivatives</i>
Orange County	2000	<i>Reverse repurchase agreements & leveraged structured notes</i>
Proctor & Gamble	157	<i>Leveraged German marks – US dollars spread</i>

Source: Exhibit 1 in McCarthy (2000), taken from Brian Kettle, Derivatives: Valuable Tool or Wild Beast?

Copyright © 1999 by Global Treasury News (www.gtnews.com).

²Details of other financial scandals can be found at <http://www.ex.ac.uk/~rdavies/arian/scandals/>.

Much more recently, improper trading in mutual funds has cost banks and some funds management companies millions of dollars in fines. For instance, in mid-March of 2004 Bank of America Corp. and its merger partner, FleetBoston Financial Corporation, agreed to pay a collective sum of \$675m to settle charges with securities regulators that they had defrauded shareholders by allowing select investors to trade improperly in their mutual funds. The Boston Globe reported on 16 March 2004: ‘With this agreement, mutual fund firms have now reached settlements totalling \$1.65 billion, eclipsing the \$1.4 billion Wall Street firms agreed to pay [in 2003] to settle charges their analysts issued biased research to win investment banking business, New York Attorney General Eliot Spitzer said.’

III.C.1.3 Defining Operational Risk

What is operational risk? This depends on what an institution wishes to gain from its operational risk management function. No two institutions will have exactly the same definition of what operational risk means since there are unique facets such as composition of the business portfolio, internal culture, risk appetite, etc., that differentiate the types of operational risks the institutions are exposed to. Nevertheless, there are some very clear commonalities that are shared by different financial institutions.

There are many highfalutin and facetious ways to define operational risk. The most important element to take into account, however, is to choose a definition that is in line with the institution’s philosophy and sound management culture of taking *proactive* stances in managing the risks of the enterprise.

One of the earliest definitions in the financial industry broadly defines operational risk in financial institutions as the ‘risk that external events, or deficiencies in internal controls or information systems, will result in an economic loss – whether the loss is anticipated to some extent or entirely unexpected’. There are two obvious observations here. This early industry definition identifies both the expected and unexpected losses attributable to operational mishaps. In simple terms, expected losses are those losses incurred during the natural course of doing business, and unexpected losses are usually associated with big surprises resulting from lapses in management and breakdown in controls. The scope of operational risk in this early definition is quite broad and extends to all facets and aspects of risk associated with both internal and external events, tangible resources such as information technology and systems, and intangibles such as people and process.

This early industry definition eventually became the cornerstone of the official definition from Basel II. The first Basel definition of operational risk was simply ‘the risk of direct or indirect loss

resulting from inadequate or failed internal processes, people and systems or external events’. This definition includes legal risk, but Basel II explicitly excluded strategic and reputational risk. These exclusions are very important aspects of the daily operation of any financial institution, but are admittedly much more difficult to assess and manage.

Concerns were expressed about the exact meaning of direct and indirect loss. Consequently the current Basel II definition drops this distinction but provides clear guidance on which losses are relevant for regulatory capital purposes. This is achieved by defining the types of loss events that should be recorded in internal loss data. In its September 2001 press release for the ‘Working Paper on the Regulatory Treatment of Operational Risk’, the Risk Management Group (RMG) of the Basel Committee on Banking Supervision defined operational risk as ‘the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events’.

According to the RMG press release this is a ‘causal-based’ definition: ‘It is important to note that this definition is based on the underlying causes of operational risk. It seeks to identify why a loss happened and at the broadest level includes the breakdown by four causes: people, processes, systems and external factors. This “causal-based” definition, and more detailed specifications of it, is particularly useful for the discipline of managing operational risk within institutions. However, for the purpose of operational risk loss quantification and the pooling of loss data across banks, it is necessary to rely on definitions that are readily measurable and comparable. Given the current state of industry practice, this has led banks and supervisors to move towards the distinction between operational risk causes, actual measurable events (which may be due to a number of causes, many of which may not be fully understood), and the P&L effects (costs) of those events. Operational risk can be analysed at each of these levels.’ The Basel II definition is primarily for capital adequacy purposes. That is, a key output of the regulatory framework is a measure of the amount of capital required by a financial institution as a buffer against unexpected operational risks.

III.C.1.4 Types of Operational Risk

The types of operational risk encountered daily within an institution are quite diverse and plentiful. What are the main types of operational risk financial institutions need to be wary of? The RMG struggled with this very same issue when it embarked on its event-by-event loss data collection exercise in June 2002.³ In its press release, the Basel Committee on Banking Supervision (2002) described the goals of this exercise: ‘The primary purpose of this survey is to

³ Having had two previous quantitative impact studies (QIS) in the previous years, the RMG decided that this more recent loss data collection exercise would specifically concentrate on very granular loss data.

collect granular (event-by-event) operational risk loss data to help the Committee determine the appropriate form and structure of the AMA (Advanced Measurement Approach). To facilitate the collection of comparable loss data at both the granular and aggregate levels across banks, the Committee is again using its detailed framework for classifying losses. In the framework, losses are classified in terms of a matrix comprising eight standard business lines and seven loss event categories. These seven event categories are then further divided into 20 sub-categories and the Committee would like to receive data on individual loss events classified at this second level of detail if available.’ The eight standard business lines are: corporate finance; trading and sales; retail banking; commercial banking; payment and settlement; agency services; asset management; and retail brokerage.

Table III.C.1.2: Basel II definition of business lines

Business Unit	Business lines		Activity Groups
	Level 1	Level 2	
INVESTMENT BANKING	Corporate Finance	Corporate Finance	Mergers and Acquisitions, Underwriting, Privatisations, Securitisation, Research, Debt (Government, High Yield) Equity, Syndications, IPO, Secondary Private Placements
		Municipal/Government Finance	
		Merchant Banking	
		Advisory Services	
	Trading & Sales	Sales	Fixed Income, equity, foreign exchanges, commodities, credit, funding, own position securities, lending and repos, brokerage, debt, prime brokerage
		Market Making	
		Proprietary Positions	
Treasury			
BANKING	Retail Banking	Retail Banking	Retail lending and deposits, banking services, trust and estates
		Private Banking	Private lending and deposits, banking services, trust and estates, investment advice
		Card Services	Merchant/Commercial/Corporate cards, private labels and retail
	Commercial Banking	Commercial Banking	Project finance, real estate, export finance, trade finance, factoring, leasing, lends, guarantees, bills of exchange
	Payment and Settlement ¹	External Clients	Payments and collections, funds transfer, clearing and settlement
	Agency Services	Custody	Escrow, Depository Receipts, Securities lending (Customers) Corporate actions
		Corporate Agency	Issuer and paying agents
Corporate Trust			
OTHERS	Asset Management	Discretionary Fund Management	Pooled, segregated, retail, institutional, closed, open, private equity
		Non-Discretionary Fund Management	Pooled, segregated, retail, institutional, closed, open
	Retail Brokerage	Retail Brokerage	Execution and full service

In Table III.C.1.2 we see that investment banking as a primary business unit is further split up into two level 1 sub-units: corporate finance; and trading and sales. Furthermore, within the trading and sales sub-unit, there are further level 2 sub-delineations: sales; market making; proprietary positions; and treasury activities. Each of these sub-units is classified based on its respective business functions, such as fixed income, foreign exchange, equity, commodities, credit, funding, brokerage, and so on.

The Committee proposes looking at seven loss event categories associated with each business unit. These are depicted in Table III.C.1.3. The level 1 event types are the practical and obvious events: internal fraud; external fraud; employment practices and workplace safety; clients, products and business practices; damage to physical assets; business disruption and system failures; and execution, delivery and process management. Furthermore, within the level 1 event type category, there are at least two level 2 sub-categories, 20 sub-categories in total. Each of these sub-categories is again classified based on the associated business activities.

Table III.C.1.3: Basel II definition of loss event types

Event-Type Category (Level 1)	Definition	Categories (Level 2)	Activity Examples (Level 3)
Internal fraud	Losses due to acts of a type intended to defraud, misappropriate property or circumvent regulations, the law or company policy, excluding diversity/discrimination events, which involves at least one internal party.	Unauthorised Activity	Transactions not reported (intentional) Tons type unauthorised (w/monetary loss) Mismarketing of position (intentional)
		Theft and Fraud	Fraud / credit fraud / worthless deposits Theft / extortion / embezzlement / robbery Misappropriation of assets Malicious destruction of assets Forgery Check kiting Smuggling Account take-over / impersonation / etc. Tax non-compliance / evasion (willful) Bribes / kickbacks Insider trading (not on firm's account)
External fraud	Losses due to acts of a type intended to defraud, misappropriate property or circumvent the law, by a third party	Theft and Fraud	Theft/Robbery Forgery Check kiting
		Systems Security	Hacking damage Theft of information (w/monetary loss)
Employment Practices and Workplace Safety	Losses arising from acts inconsistent with employment, health or safety laws or agreements, from payment of personal injury claims, or from diversity / discrimination events	Employee Relations	Compensation, benefit, termination issues Organised labour activity
		Safe Environment	General liability (slip and fall, etc.) Employee health & safety rules events Workers compensation
		Diversity & Discrimination	All discrimination types
Clients, Products & Business Practices	Losses arising from an unintentional or negligent failure to meet a professional obligation to specific clients (including fiduciary and suitability requirements), or from the nature or design of a product.	Suitability, Disclosure & Fiduciary	Fiduciary breaches / guideline violations Suitability / disclosure issues (KYC, etc.) Retail consumer disclosure violations Breach of privacy Aggressive sales Account churning Misuse of confidential information Lender Liability
		Improper Business or Market Practices	Antitrust Improper trade / market practices Market manipulation Insider trading (on firm's account) Unlicensed activity Money laundering
		Product Flaws	Product defects (unauthorised, etc.) Model errors

Table III.C.1.3 (Continued): Basel II definition of loss event types

Event-Type Category (Level 1)	Definition	Categories (Level 2)	Activity Examples (Level 3)
		Selection, Sponsorship & Exposure	Failure to investigate client per guidelines Exceeding client exposure limits
		Advisory Activities	Disputes over performance of advisory activities
Damage to Physical Assets	Losses arising from loss or damage to physical assets from natural disaster or other events.	Disasters and other events	Natural disaster losses Human losses from external sources (terrorism, vandalism)
Business disruption and system failures	Losses arising from disruption of business or system failures	Systems	Hardware Software Telecommunications Utility outage / disruptions
Execution, Delivery & Process Management	Losses from failed transaction processing or process management, from relations with trade counterparties and vendors	Transaction Capture, Execution & Maintenance	Miscommunication Data entry, maintenance or loading error Missed deadline or responsibility Model / system misoperation Accounting error / entity attribution error Other task misperformance Delivery failure Collateral management failure Reference Data Maintenance
		Monitoring and Reporting	Failed mandatory reporting obligation Inaccurate external report (loss incurred)
		Customer Intake and Documentation	Client permissions / disclaimers missing Legal documents missing / incomplete
		Customer / Client Account Management	Unapproved access given to accounts Incorrect client records (loss incurred) Negligent loss or damage of client assets
		Trade Counterparties	Non-client counterparty misperformance Misc. non-client counterparty disputes
		Vendors & Suppliers	Outsourcing Vendor disputes

For internal purposes it is very important to establish suitable definitions of the different types of operational risk that are relevant for each individual institution. Once suitable definitions of event types are decided upon, the institution can proceed to establish a structure for the operational risk management framework. It is interesting to note that almost all of the so-called internationally active banks that are the primary focus of Basel II have their own unique definitions of operational risk event types that are tailored to their particular businesses, corporate culture and risk appetite. For example, one internationally active bank defines operational risk as: ‘The risk of inadequate identification of and/or response to shortcomings in organizational structure, systems, transaction processing, external threats, internal controls, security measures and/or human error, negatively affecting the bank’s ability to realize its objectives.’ A smaller domestic bank defines operational risk simply as: ‘The risk associated with the potential for systems failure in a given market’.

III.C.1.5 Aims and Scope of Operational Risk Management

The fundamental goal of operational risk management should be risk *prevention*. The assessment (meaning the quantitative measurement) of operational risk is of secondary importance.

Because complete elimination of operational risk failures is not feasible, our operational risk management framework must aim to *minimise the potential for loss* – through whatever means possible. Indeed, the risk management of the *entire* institution as an enterprise should focus more on the *operational* aspects of the different business activities – with the important provision that

the other key risk management functions within the enterprise (e.g., market risk, credit risk, audit, and compliance) are already firmly grounded.

Regardless of how an institution chooses to define operational risk, it is vitally important at the outset to explicitly articulate what its target objectives and key concerns are. I suggest the following important objectives when establishing an operational risk management function:

1. To formally and explicitly *define and explain* what the words ‘operational risk’ mean to the institution.
2. To avoid potential catastrophic losses.
3. To enable the institution to anticipate all kinds of risks more effectively, thereby preventing failures from happening.
4. To generate a broader understanding of enterprise-wide operational risk issues at all levels and business units of the institution – *in addition* to the more commonly monitored credit risk and market risk.
5. To make the institution less vulnerable to such breakdowns in internal controls and corporate governance as fraud, error, or failure to perform in a timely manner which could cause the interests of the institution to be unduly compromised.
6. To identify problem areas in the institution *before* they become critical.
7. To *prevent* operational mishaps from occurring.
8. To establish *clarity* of people’s roles, responsibilities and accountability.
9. To strengthen management oversight at all levels.
10. To identify business units in the institution with high volumes, high turnover (i.e., transactions per unit time), high degree of structural change, and highly complex support systems. Such business units are especially susceptible to operational risk.
11. To empower business units with the *responsibility* and *accountability* of the business risks they assume on a daily basis.
12. To provide objective measurements of performance for operational risk management.
13. To monitor the danger signs of both income *and* expense volatilities.
14. To effect a change of behaviour within the institution and to enhance the culture of control and compliance within the enterprise.
15. To ensure that there is compliance to all risk policies of the institution and to modify risk policies where appropriate.
16. To provide objective information so that *all* services offered by the institution take account of operational risks.
17. To ensure that there is a clear, orderly and concise measure of due diligence on all risk-taking and non-risk-taking activities of the institution.

18. To provide the executive committee⁴ regularly with a concise ‘state of the enterprise’ report for strategic and planning purposes.

The stated objectives tacitly assume that the institution already has in place robust credit risk and market risk management functions, supported by audit, compliance and risk control oversight. Note that the operational risk management objectives delineated above should apply to all business units, including those responsible for market risk and credit risk management.

In practice, the scope of an operational risk management function within an institution should aim to encompass virtually *any* aspect of the business process undertaken by the enterprise. The scope must transcend those business activities that are traditionally most susceptible to ‘operations risk’ – that is, those activities with high volume, high turnover, and highly complex support systems, e.g. trading units, back office, and payment systems. It is true that the business activities sharing these characteristics have the greatest exposure to operational risk failures. Nevertheless, other business activities could potentially sustain economic losses of similar magnitude.

Table III.C.1.4: Two broad categories of operational risk

Operational strategic risk (‘external’)	Operational failure risk (‘internal’)
<p data-bbox="277 1133 746 1196"><i>Defined as the risk of choosing an inappropriate strategy in response to external factors such as:</i></p> <ul style="list-style-type: none"> <li data-bbox="277 1234 427 1263">• political <li data-bbox="277 1272 427 1301">• taxation <li data-bbox="277 1310 448 1339">• regulation <li data-bbox="277 1348 427 1377">• societal <li data-bbox="277 1386 475 1415">• competition 	<p data-bbox="836 1133 1326 1196"><i>Defined as the risk encountered in the pursuit of a particular chosen strategy due to:</i></p> <ul style="list-style-type: none"> <li data-bbox="858 1234 986 1263">• people <li data-bbox="858 1272 991 1301">• process <li data-bbox="858 1310 1034 1339">• technology <li data-bbox="858 1348 975 1377">• others

Source: adapted from Cronby et al. (1998)

Should operational risk management encompass external events? The goal of operational risk management must be to focus on internal processes (as opposed to external events) since only internal processes are within the control of the firm. The firm’s *response* to external events is, however, a valid concern for operational risk management. Hence we examine operational risk from two interrelated perspectives. Table III.C.1.4 distinguishes between operational ‘strategic’ risk (i.e., a flawed *internal* response to *external* stimuli) and operational ‘failure’ risk. Failure to

⁴ In this context, the Executive Committee, composed only of very senior members of management, is assumed to be the highest governing body of the institution.

comply with externally dictated strategic risk factors – such as changes in tax laws⁵ and new derivatives accounting treatment (e.g., FASB 133) – ultimately translates to an internal operational risk failure. Once senior management issues the call to action in response to an external stimulus, there must be no *internal* breakdown in the people, processes and technologies supporting the strategic call to action.

III.C.1.6 Key Components of Operational Risk

In view of the very wide scope of an institution's operational risk management function, we might want to concentrate on some key components of operational risk, such as the following.

(i) Core operational capability

Risks to the institution's core operational capability include the risk of premises, people or systems becoming unavailable due to: natural disasters, fire, bombs or technical glitches; loss of utilities such as power, water or transportation; employee disputes such as strikes; loss of key operational personnel; and the loss or inadequacy of systems capabilities due to computer viruses or Y2K issues. All of the aforementioned events seriously disrupt the institution's core competency in supporting its *long-term* and *stable* operations, thereby representing a considerable exposure in terms of their possible impact on the institution's future earnings and credibility. The good news is, for the most part, that many of the risks mentioned above are largely *insurable* at some cost to the institution. Insurance is, therefore, a useful risk mitigant for these kinds of failure risk.⁶

(ii) People⁷

An institution's most important assets are its good people. Unfortunately, people also contribute a myriad of problems through: human error; fraud,⁸ lack of honesty⁹ and integrity; lack of cooperation and teamwork; in-fighting, jealousies, and rumour-mongering; personal sabotage;

⁵ For example, an institution operating in another country might encounter some unexpected tax liability due to an unforeseen change in local taxation rules that was not anticipated by the accounting department. This is definitely an operational risk item that could lead to large unanticipated fines and tax liabilities.

⁶ Basel II currently does not recognise insurance as a risk mitigant, except in some restricted cases within the Advanced Measurement Approach. This is somewhat odd considering the fact that banks have routinely used insurance as a risk management tool.

⁷ Unfortunately, the category of *people* is by far the largest cause of operational risk failures.

⁸ An FDIC study found that fraud was the main contributing factor to 25% of 92 bank failures in the period 1960–77. The proportion rises to 83.9% if one includes 'insider fraud' – i.e., improper lending to individuals or groups connected with the bank. A review by the Bank of England suggested that, in the UK in the period 1984–96, fraudulent concealment was a major contributory factor in 7 out of 22 cases of bank problems. In many of these cases, these bank frauds were perpetrated by senior managers of the banks themselves.

⁹ On the subject of integrity and honesty, 'rogue' trading generally surfaces as the most obvious case of dishonesty and fraud; however, we need to recognise that the problem of fraud also extends to teller theft, mailroom theft, illegal funds transfers, and 'insider' fraud mentioned in footnote 8, etc. Some well-known cases of securities fraud and rogue trading are: Kidder-Peabody (Jett, \$340m, 1994), Orange County (Citron, \$1.6bn, 1994), Barings (Leeson, \$1.2bn,

office politics;¹⁰ lack of segregation and risk of collaboration; lack of professionalism and customer focus; over-reliance on key individuals, insufficient skills, training and education; insubordination; employment disputes; poor management and supervision; and a lack of culture of control, discipline and compliance.¹¹

(iii) Client relationships

An institution derives much of its value from its reputation and the services it provides to its client base. Any damage to an institution's reputation has the potential to disrupt revenue flow. From an operational risk perspective, the institution needs to assess how disreputable activities might harm its client relationships. Examples include: money laundering; Nazi gold;¹² improper client suitability and lack of disclosure;¹³ false valuations of client assets to mislead or conceal losses; collusional relationships with broker-dealers; cosy association with highly-leveraged institutions; and dishonest practices amidst competition that can harm the institution's reputation.

(iv) Transactional and booking systems

Operational risk failures are no longer limited to settlement risk in the trading accounts or back office. More recently, with advances in automation, transactional issues also include: data capture and processing; deal confirmation¹⁴ and contractual documentation, e.g., ISDA master agreements; collateral management; and general processing and payment/settlement errors which not only disrupt the flow of business but also put an institution at risk of litigation. In addition, *corporate banking* activities contributing to operational failures may include: correspondent banking; payment services; treasury services; private trust and executor services; structured finance; custody; and leasing. From a *retail banking* perspective, there is even more room for potential operational failures associated with such retail banking activities as: mortgage servicing; funds management; deposit taking; sending; foreign exchange; custody; credit cards; ATMs; private banking; and insurance. The transactional processes associated with handling retail customers are many: payments; cheque clearing; cash handling and teller errors; credit analysis; account opening; documentation; mortgage applications; interest charges; processing credit/debit card transactions; processing insurance claims; and payroll processes. In addition, associated with each of these

1995), Daiwa (Iguchi, \$1.1bn, 1995), Sumitomo (Hamanaka, \$1.8bn, 1995) and many other cases involving varying amounts of losses.

¹⁰ Ask yourself this question: how many institution-wide problems and inefficiencies were caused by internal fights and office politics?

¹¹ Rogue trading is not the only contributor to well-publicised derivatives losses. Table III.C.1.1 is a list of the largest derivatives losses attributable largely to failures in risk management where operational risk played a major role.

¹² This has become an important reputation risk issue among a few big European banks in the recent past.

¹³ Among the most highly publicised client suitability and disclosure cases are the Gibson Greetings and Procter & Gamble lawsuits against Banker's Trust (1994) and the Orange County collective legal debacle with Merrill Lynch, Morgan Stanley Dean Witter, and Nomura Securities (1994).

¹⁴ The best-known documented case of booking errors occurred at Salomon Brothers in the mid-1990s where a back-office confirmation for a trade was erroneously booked several orders of magnitude larger than the intended trade.

business processes is heavy reliance on a sound and stable systems infrastructure within the institution.

(v) Reconciliation and accounting

Of course the reconciliation of transactions at different levels of institutional activities is important from a bookkeeping perspective. Another important aspect of reconciliation and accounting is that it enables the institution to identify areas of inefficient capital allocation. More broadly, a finance or accounting department, in a quagmire of bureaucracy and mere paper-pushing, can do the institution a lot of harm by failing to provide senior management with a precise picture of the state of the institution's finances. The inability to reconcile properly the revenue-generating activities with the general ledger *per se* inhibits the institution from strategically assessing the performances of its business units and their growth potential.¹⁵ In addition, it also undermines the institution's ability properly to allocate its scarce resources, such as capital. Finally, the inability of the finance, legal or accounting departments fully to assess the implications of regulatory changes puts the institution at a serious disadvantage with regard to tax shelters and favourable legal treatment of the institution's assets and liabilities.

(vi) Change and new activities

To stagnate is to fall behind. But, in responding to rapid industry developments, the institution should not stumble and fall when initiating new business activities. For example, the introduction of the euro or the change in regulatory accounting rules, e.g., FAS 133,¹⁶ requires the institution to be more vigilant in implementing new technology, re-engineering its processes, expanding its staff, accepting new clients, launching new products or entering into new markets. More recently, we have new regulatory directives for anti-money laundering and terrorist funding, the Sarbanes-Oxley Act of 2002 which requires all listed institutions to enforce more effective corporate governance and more effective financial statements reporting, and many other new directives concerning securities fraud promoted by the Securities and Exchange Commission. Inability to adapt to change may damage an institution's reputation or disrupt the continuity of its old businesses.¹⁷

¹⁵ Ask yourself this question: through our current finance and accounting systems, do we know for sure where we are generating the greatest revenue for the least amount of risks that we take?

¹⁶ Ask yourself this question: how quickly can the institution conform to the FAS 133 directives without unduly taxing its resources and disrupting its day-to-day business operations?

¹⁷ Ask yourself this question: can the institution leverage its current resources and technology and enter into a new market without unduly incurring additional expenses? If the answer is negative, there is operational risk in the headline risk categories involving people, processes and technology.

(vii) *Expense and revenue volatility*

A sure sign of operational failures associated with management control is a rapid increase in expenses and significant deviations from budget. Expense increases (including bonuses, salaries, and systems infrastructure spending) *without adequate return* signals a potential breakdown and inefficiencies in people, process and technology. Rapid increase in expenses is not necessarily a desirable symbol of growth but, in my years of observation, it is also a sure sign of lax accounting and a complacent management on the verge of going out of control.¹⁸ It is normally associated with a bank's undesirable corporate culture of wanton waste and lack of accountability. On a related matter, excessive revenue volatility is the result of at least two related factors: the inability to respond properly to external market conditions; and the failure to control the operational cost base. Each of these key factors has its *obvious* attendant operational failures in people, process and technology in responding to external risk factors. I need not elaborate further on this.

III.C.1.7 Supervisory Guidance on Operational Risk

The Basel Committee on Banking Supervision has provided preliminary supervisory guidelines for the management and supervision of operational risk. The guidelines are intended to serve as best or *sound practices* within the financial industry. The Committee 'recognises that the exact approach for operational risk management chosen by an individual bank will depend on a range of factors, including its size and sophistication and the nature and complexity of its activities. However, despite these differences, clear strategies and oversight by the board of directors and senior management, a strong operational risk culture and internal control culture (including, among other things, clear lines of responsibility and segregation of duties), effective internal reporting, and contingency planning are all crucial elements of an effective operational risk management framework for banks of any size and scope. The Committee therefore believes that the principles outlined in this paper establish sound practices relevant to all banks' (Basel Committee on Banking Supervision, 2003).

The Committee also recognizes that '*internal operational risk culture* is taken to mean the combined set of individual and corporate values, attitudes, competencies and behaviours that determine a firm's commitment to and style of operational risk management.' Recognising the different nature of operational risk in different institutions, the Committee defines the *management* of operational risk to mean the 'identification, assessment, monitoring and control/mitigation' of risk. To this end, the sound practice paper of February 2003 is structured around *ten basic principles* grouped into four main themes:

¹⁸ Ask yourself this question: in the past three years, did the increase in expenditure result in a greater market share or revenue to the institution? If the answer is negative, there is operational risk involving inefficiency and misallocation of precious resources.

- Developing an appropriate risk management environment.
- Risk management: identification, assessment, monitoring, and mitigation/control.
- Role of supervisors.
- Role of disclosure.

By imposing the ultimate responsibility on senior management via the board of directors, the first theme emphasizes the importance of cultivating a risk-awareness culture within the institution as dictated directly from the highest level of the organization – the board of directors. The second theme maintains that the management of risk is comprised of four important complementary activities. The first concerns itself with surveillance and identification of risk within the institution, followed by a thorough assessment and monitoring of events as they unfold, and finally, devising mechanisms to control and mitigate these risks, even before they occur. As stated earlier, prevention is key. Recognising that the safety and soundness of the financial system is a collaborative effort, the third theme emphasises the important role regulatory supervisors play in the risk management process of financial institutions. Finally, risk management can only be facilitated properly if the process is clear and transparent at the outset. Public disclosure to the market, therefore, serves as an important binding constraint on the behaviour of financial institutions as they provide the public with their intermediary functions.

III.C.1.8 Identifying Operational Risk – the Risk Catalogue

How can an institution identify potential operational risks lurking within the organization? In the recent past, many institutions have been surprised to discover that even the most obvious types of operational risk are widely prevalent within the organisation, unnecessarily squandering precious resources and hindering productivity.

The first step in the identification process is to require that each *business unit* (or *operational unit*) shall be assessed using a so-called *risk catalogue* which adequately identifies all the risk categories relevant to the specific unit being assessed. It clearly identifies possible operational failures under three categories: people, process and technology.

Table III.C.1.5: Risk catalogue for Business Unit A¹⁹

People Risk			
Incompetency	<input checked="" type="checkbox"/>	Internal Politics,	<input checked="" type="checkbox"/>
Inadequate Head Counts	<input checked="" type="checkbox"/>	Conflict of Interest,	<input checked="" type="checkbox"/>
Key Personnel Management	<input checked="" type="checkbox"/>	Lack of Cooperation	<input checked="" type="checkbox"/>
Communication	<input checked="" type="checkbox"/>	Collusion and Connivance	<input checked="" type="checkbox"/>
		Fraud	
Process Risk			
A. Model Risk			
Model or Methodology Error	<input checked="" type="checkbox"/>		
Pricing or Mark-to-Model Error	<input type="checkbox"/>		
Availability of Loss Reserves	<input type="checkbox"/>		
Model Complexity	<input type="checkbox"/>		
B. Transaction Risk			
Execution Error	<input checked="" type="checkbox"/>	Capacity Risk	<input checked="" type="checkbox"/>
Booking Error	<input checked="" type="checkbox"/>	Valuation Risk	<input checked="" type="checkbox"/>
Collateral, Confirmation, Matching, and Netting Error	<input checked="" type="checkbox"/>	Erroneous Disclosure Risk	<input type="checkbox"/>
Product Complexity	<input type="checkbox"/>	Fraud	<input type="checkbox"/>
C. Operations Control Risk			
Limit Exceedances	<input checked="" type="checkbox"/>		
Volume Risk	<input checked="" type="checkbox"/>		
Security Risk	<input type="checkbox"/>		
Position Reporting Risk	<input type="checkbox"/>		
Profit and Loss Reporting Risk	<input checked="" type="checkbox"/>		
Technology Risk			
Systems Failure	<input type="checkbox"/>	Programming Error	<input checked="" type="checkbox"/>
Network Failure	<input checked="" type="checkbox"/>	Data Corruption	<input checked="" type="checkbox"/>
Systems Inadequacy	<input checked="" type="checkbox"/>	Disaster Recovery Risk	<input checked="" type="checkbox"/>
Compatibility Risk	<input type="checkbox"/>	Systems Age	<input checked="" type="checkbox"/>
Supplier/Vendor Risk	<input checked="" type="checkbox"/>	Systems Support	<input checked="" type="checkbox"/>

Consider a simple example: suppose we have determined that Business Unit A, due to its intrinsic business activities, is subject to some sources of operational risk. We can then check them off against our generic risk catalogue as shown in Table III.C.1.5. This checklist can then form the basis for assessing the loss *frequency* and loss *severity* of the different event types in the business unit using the *risk scorecard* (see next section). We shall also see, in Section III.C.1.10, that the control process requires the identification of pertinent operational risk failures at two different levels: *independent management oversight* and *self-assessments by individual business units*.

III.C.1.9 The Operational Risk Assessment Process

It is important to keep in mind that an operational risk assessment process without the aim of proactive management is an exercise in futility. What drives our desire to ‘measure’ must be our belief that a sound and active risk management structure is in place – or will be in place in the future. As a precursor to building a sound risk management structure, measurement can be the tool by which senior managers are convinced that such a structure is needed. It also helps an institution to identify the key problem areas of operational risk and this helps target the resources that will be allocated.

¹⁹ Risk types shown are for illustration only. In practice, different business units may have different types of risks they are most concerned with. For instance, there is presumably no model risk in retail banking or in leasing.

For each business unit, the risk assessment process follows four fundamental steps: (1) inputs to risk catalogue; (2) risk assessment scorecard; (3) review and validation; and (4) outputs of risk assessment process.

Step 1: Inputs to Risk Catalogue

Operational risk should be evaluated *net* of risk mitigants. For example, if the institution has insurance to cover a potential breakdown, then the degree of risk must be properly adjusted by the insurance premium paid. To obtain a measure of *net* operational risk, the required inputs to the risk catalogue must be able to adequately assess both the *frequency* of failure occurrences and the *severity* of loss given that a failure occurs:

- The assessment for frequency of occurrences may come from both internal and external reports, such as: audit reports; external audit reports; regulatory reports; management reports; expense reports; deviation from business plans, operational plans, and budgets, etc.; and expert opinion and industry ‘best practices’.
- An assessment for severity of loss may come from: management interviews, both *pre* and *post mortem*; variances on budgets; insurance claims; and loss history, whenever possible.²⁰

Step 2: Risk Assessment Scorecard

Using the risk catalogue and the inputs from step 1, each business or operational unit will be assessed using a *risk scorecard*. The risk scorecard will appropriately identify and assess the nature of operational risk based on the following broad points:

- *Risk categories* – people, process, technology, and external dependencies.
- *Connectivity and interdependencies*. Because the headline risk categories of people, process and technology cannot be looked at in isolation, their cumulative effects and *interdependencies* must be carefully identified and accounted for.
- *Change, complexity, and complacency*. The sources that drive the headline risk categories may be due to: a *change* in the work environment, e.g., the introduction of new technology to the business unit; the *complexity* of products, process or technology; or the *complacency* factor due to ineffective management of the unit.

²⁰ A few major banks are beginning to gather their own internal loss experiences – the outcome will not be known until many years from now. Loss history should also cover credit losses as a result of operational mishaps, loss due to theft and fraud, and losses strictly due to errors. Admittedly, this is an extremely difficult task and the financial industry is still struggling with how to collect these loss data. In addition, the RMG has also analysed data collected from the numerous participating banks through the 2002 Operational Risk Loss Data Collection Exercise (LDCE) in June of 2002. The 2002 LDCE was an extension and refinement of two previous data collection exercises sponsored

- *Frequency and severity assessments.* Quantifying the likelihood of breakdown in operational processes is very difficult. It may be simply ‘rated’ as very likely, not likely, very unlikely and so forth, or a question relating to the expected number of loss events may be posed. Severity of loss describes the potential monetary loss to the institution, given the occurrence of an operational failure. Since actual loss history may be difficult to come by, some institutions *subjectively* attach a range of loss (e.g., between \$5 million to \$10 million for certain failures). More details on the recommendations for frequency and severity of self-assessments are given in Chapter III.C.3.
- *Net operational risk.* Operational risks should be evaluated net of risk mitigants. For instance, the potential monetary amount lost due to certain insurable operational failures can be reduced through the use of risk mitigants, e.g., insurance and underwriting. We need to find out which bank activities are currently covered by insurance policies and by how much. In addition, a catalogue of insurable bank activities needs to be prepared.
- *Net risk assessment.* The combination of all the ingredients in the risk scorecard enumerated above gives the overall net risk assessment.

Step 3: Review and Validation

After the risk assessment process is completed (via the risk catalogues) and risk scorecards for *each* business unit are produced, it is the responsibility of the *operational risk management committee*²¹ to review the assessment results with the management of the respective business unit and other key officers of the institution. The responsibilities of the committee may include:

- Formulating a set of operational risk policies and guidelines clearly delineating the actions needed to correct and prevent the operational problems and issues identified.
- Determining the important differences between the unit's own self-assessment and the independent assessment.
- Opining on the ratings in the risk scorecards before publication.
- In conjunction with audit and compliance departments, issuing a mandatory report and list of recommendations to the affected business units.
- Issuing summary risk reporting about the enterprise to the executive committee.

Step 4: Outputs of Risk Assessment Process

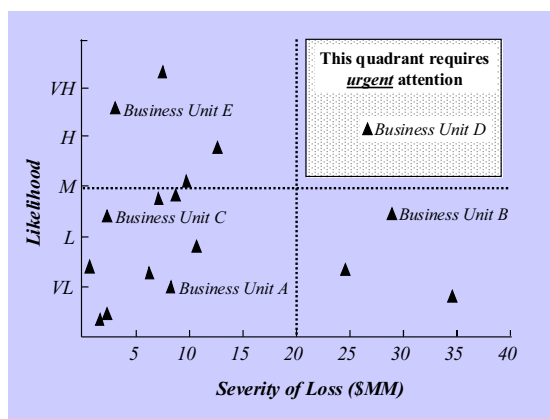
There are several possible outputs from the operational risk assessment process. We concentrate on only three broad items.

by the RMG, which primarily ‘focused on banks’ internal capital allocations for operational risk and their overall operation risk loss experience during the period from 1998 to 2000’.

(i) *Improved Risk Reporting and Analysis:*

As an ongoing goal of the operational risk assessment framework, the institution should endeavour to streamline its risk reporting processes among the different risk-monitoring units of the institution (i.e., audit, compliance, and risk control). These reports should be viewed as a concise summary of specific audit and compliance reports which are already instituted within the financial institution. The most useful of these reporting tools are the risk catalogue, risk scorecards and ‘heat maps’ which are used to highlight the relative information on operational risk exposures across the institution. An example of a heat map is given in Figure III.C.1.6. This shows that Business Unit D, relative to all the other business units in the institution, has a moderately high likelihood of incurring a large amount of loss due to operational risk failures. This means that it requires a relatively large amount of economic capital to sustain its business activities.

Figure III.C.1.6: Example of a heat map

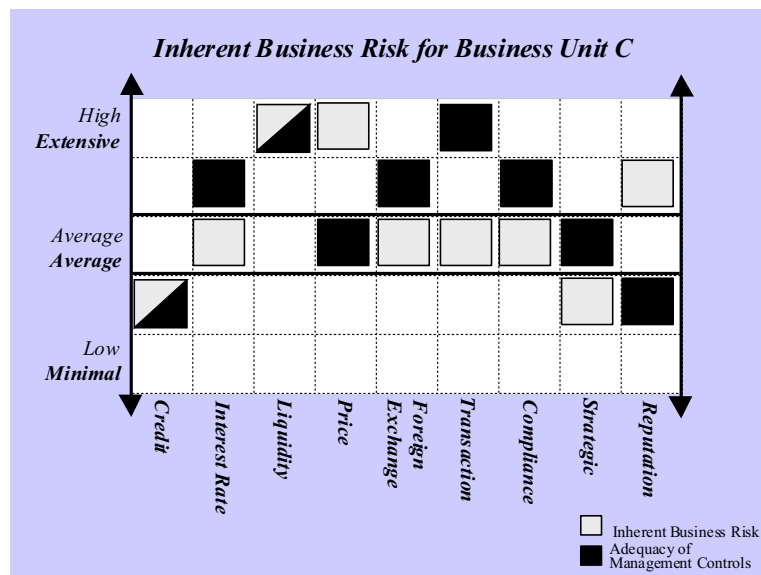


²¹ In many financial institutions, the operational risk management committee is a committee within either the risk management function or the audit function.

(ii) OCC Exam Chart

The Office of the Comptroller of the Currency (OCC), in its September 1995 press release, highlighted its examination procedure of banks to cover nine principal categories of risk. They are: credit risk; interest-rate risk; liquidity risk; price risk; foreign exchange risk; transaction risk; compliance risk; strategic risk; and reputation risk. (The OCC mandate is directed at banks with financial derivatives activities, therefore the nine categories of risk need not apply to every business units of the financial institution.) Using the risk scorecard and other reports, we can graphically represent the evaluation of a business unit in a concise manner in line with the OCC examination procedure. This is illustrated by the OCC exam chart in Figure III.C.1.7.

Figure III.C.1.7: OCC exam chart



(iii) Capital Attribution

By attributing economic capital to operational risks we can ensure that business units which are more prone to operational failures are assigned a greater allocation of capital commensurate to the risks that they take. Admittedly, this is a difficult and subjective task and a whole chapter of this handbook is devoted to it (see Chapter III.0). It is also important to note that the current Basel II proposals call for a regulatory capital charge for operational risk, in addition to the capital charge already required for both market risk and credit risk (see Chapter III.C.3). Whether there is wisdom behind a capital charge for operational risk remains to be seen (Ong, 1998). After all, major financial catastrophes resulting from breakdowns in people, process and technology cannot be prevented entirely. Thus, operational risk events cannot be eliminated altogether, regardless of the amount of capital charge levied on operational risk.

III.C.1.10 The Operational Risk Control Process

With hindsight the well-publicised derivatives losses listed in Section III.C.1.2 were all *preventable*. Outside of derivatives activities, losses attributable to fraud in other lines of business are likewise preventable. Subsequently, countless studies have continued to point to the following failures in risk control as the ultimate culprits.

- (i) *Lax management structure*: lack of adequate management oversight and accountability; no segregation of duties; too many delays in systems development; disrespect for audit reports; and ignorance.
- (ii) *Inadequate assessment of risk: on and off-balance sheet activities*: lack of stress-testing for unexpected market moves; inappropriate setting of limits; too much risk relative to capital; and lack of risk-adjusted return measurement.
- (iii) *Lack of transparency*: inaccurate information on capital, solvency and liquidity; inadequate accounting policies; lack of benchmarks and comparability; lack of approvals, verifications and reconciliations; and lack of review of operating performance.
- (iv) *Inadequate communication of information between levels of management*: lack of escalation process in times of crises; paying too much ‘lip service’ in the various risk committees; lack of procedures for monitoring and correcting deficiencies; and poor communication between different risk-monitoring groups.
- (v) *Inadequate or ineffective audit and compliance programmes*.

To help facilitate the identification of operational risk failures within the institution by business unit, it is important that the operational risk management function is streamlined alongside the risk control, compliance, and audit functions of the institution. Lessons that we have learned from many highly publicised financial fiascos all point to the need for the following:

- (i) *Independent management oversight*. This includes audit oversight, risk control and compliance functions, and most definitely senior management involvement. Each overseer plays the role of *independent* ‘risk monitor’, considering such operational performance measures as volume, turnover, settlement failures, delays, errors, compliance to market and credit limits, income and expense volatility, effectiveness of line management, accounting anomalies, and other higher-level controls. For many business activities of the holding company, this information is already available within the institution.
- (ii) *Self-assessments by individual business units*. Line management has the best knowledge of its own people, the day-to-day processes it has to go through, the integrity of the

systems supporting the business unit, and the external circumstances that could cause its people, process and technology to fail. A self-assessment by the individual business units is, therefore, a key first step in the operational risk assessment process.

III.C.1.11 Some Final Thoughts

Operational risk failures can wreak havoc within an organization if not properly identified, assessed, monitored, controlled and mitigated. Furthermore, if not sufficiently contained, operational risk also has a tendency to spill over and cause systemic risk to the broader markets. In spite of its importance for containing operational risk, it is still more art than science. Operational risk management continues to be one of the least developed areas of enterprise risk management in spite of the heightened attention it has received. Perhaps since operational risk is fundamentally qualitative in nature, it might never be as developed as other risk areas.

While much progress has been made over the past several years, at least primarily within the financial industry, the fundamental operational risk management framework continues to be confused by many people. Many people with quantitative background have used the Basel II proposals as their impetus for furthering the argument for more operational risk modelling. People with compliance, audit and risk control backgrounds tend to interpret the Basel II guidance as an opportunity to codify additional policies, thereby reducing operational risk management to a mere set of rules and regulations. In 2006 there will be a new regulatory capital charge for operational risk, but no regulators in their right mind would think that operational risk management is about levying capital charges. While capital charges are important, they are only one small component of prudent risk management, and they are not a good substitute for sound judgement. ‘An increase in capital will not itself reduce risk; only management action can achieve that’, a Moody’s Special Comment reported (Moody’s Investor’s Service, 2003).

My personal experience as head of enterprise risk management and chief risk officer for two of the ten largest banks in the world has taught me that most operational risk failures are preventable. The processes outlined in this chapter are based on my experience of how to prevent operational risk mishaps. Experience tells me that the most important aspect of the operational risk management framework is still sound corporate governance and proactive senior management involvement. After all, the control of operational risk is concerned fundamentally with good management. In practice, good management means vigilance, patience, and persistence in improving the risk management process. And this is what operational risk management is all about.

References

Basel Committee on Banking Supervision (2002) *Operational Risk Loss Data Collection Exercise – 2002*, 4 June.

Basel Committee on Banking Supervision (2003) *Sound Practices for the Management and Supervision of Operational Risk*, February.

Crouhy, M, Galai, D and Mark, R (1998) Key steps in building consistent operational risk measurement and management. In *Operational Risk and Financial Institutions*. London: Risk Books.

Financial Services Authority (2003) *Building a Framework for Operational Risk Management: The FSA's Observations*, July.

McCarthy, E. (2000) Derivatives revisited. *Journal of Accountancy*, 189(5). See <http://www.aicpa.org/pubs/jofa/may2000/mccarthy.htm>

Moody's Investors Service (2003) *Moody's Analytical Framework for Operational Risk Management of Banks*. Special Comment, January.

Ong, M (1998) On the quantification of operational risk – a short polemic. In *Operational Risk and Financial Institutions*. London: Risk Books.

Ramadurai, K., Olseon, K, Andrews, D, Scott, G., and Beck, T. (2004) *The Oldest Tale but the Newest Story: Operational Risk*. Special Report, FitchRatings, January.

III.C.2 Operational Risk Process Models

James Lam¹

III.C.2.1 Introduction

Management and board attention to operational risk management (ORM) has never been greater. While businesses have always faced operational risks, the discipline of ORM is still in the early stages of development. The focus on operational risk has been driven by a number of important factors:

- *Corporate disasters.* The need for ORM first gained the attention of risk management professionals in the 1990s when they realized that the root causes underlying the major financial disasters – Barings, Kidder, Daiwa, etc. – were operational risks and not financial risks. More recent corporate failures such as Enron and WorldCom, as well as the market-timing and late-trading problems plaguing the mutual fund industry, have reinforced the importance of ORM. In the aftermath of these disasters, the standards for corporate governance and risk management have increased. These new standards impact not only corporate executives and boards, but also key stakeholders such as stock analysts, rating agencies, and regulators.
- *Regulatory actions.* In response to the corporate disasters, regulators have dramatically increased their examination and enforcement standards. New regulations with significant operational risk requirements include Sarbanes-Oxley (in particular, Section 302 on certification of chief executives and chief financial officers and Section 404 on internal controls), the Patriot Act, anti-money laundering and bank secrecy acts, and other corporate governance rules adopted by the stock exchanges. Additionally, the new Basel initiative (Basel II) has established a direct linkage between minimum regulatory capital and a bank's underlying risks, including explicit treatment for operational risk.
- *Industry initiatives.* A number of industry initiatives have been organized around the world to establish frameworks and standards for corporate governance and risk management. These initiatives include the Treadway Report (United States, 1993) that produced the Committee of Sponsoring Organizations (COSO) framework of internal controls, while the Turnbull Report (United Kingdom, 1999) and the Dey Report (Canada, 1994) developed similar

¹ President, James Lam & Associates; founding member, Blue Ribbon Panel, PRMIA; and Senior Research Fellow, Beijing University.

guidelines. It is noteworthy that the Turnbull and Dey reports were supported by the stock exchanges in London and Toronto, respectively. In 2004, COSO is scheduled to release a major study on enterprise-wide risk management (ERM), which will include key ERM principles and advocate its application within a sound corporate governance framework.

- *Corporate programs.* Corporations have achieved significant benefits from their risk management programs; among these are stock price improvement, debt rating upgrades, early warning of risks, loss reduction, and regulatory capital relief. While ORM programs are relatively new, early adapters have reported sustained reduction in operational losses and error rates (one company reported a sustained 80% reduction in operational risk losses). Other reported benefits include improved customer service and operational efficiency. These results demonstrate that investments in operational risk controls can produce direct benefits that are multiples of the costs, as well as indirect benefits such as prevention of crises that divert management attention and cause reputational damage.
- *Technology developments.* Over the past decade, technology developments have transformed how businesses operate. Examples include using the Internet to communicate with customers and facilitate commerce; developing customer relationship management applications to better serve customer segments; and outsourcing IT operations and business processes to improve efficiency. In risk-intensive industries, such as financial and energy services, corporations have also developed sophisticated models and databases to measure all types of risk. While these technology developments provide business benefits, they also present new and complex risks, such as information security, data integrity, cyber-crime, cyber-terrorism, systems availability, and model risk. These risks require operational risk controls for day-to-day operations, as well as disaster recovery planning for unlikely but potentially disastrous events.

Going forward, the key trends and developments highlighted above should continue to assert significant pressure on corporate boards and executives to improve their risk management capabilities, especially in the area of operational risk.

The focus of this chapter is on the development and application of operational risk process models. We will discuss the following questions:

- How to develop and apply operational risk process models?
What are the specific quantitative and qualitative tools used by companies today?
- How to link these tools with economic capital allocation?
- What are the actions management can take to mitigate operational risk?

At the end of this chapter, we will use IT outsourcing as an example to illustrate how an operational risk process can be established.

III.C.2.2 The Overall Process

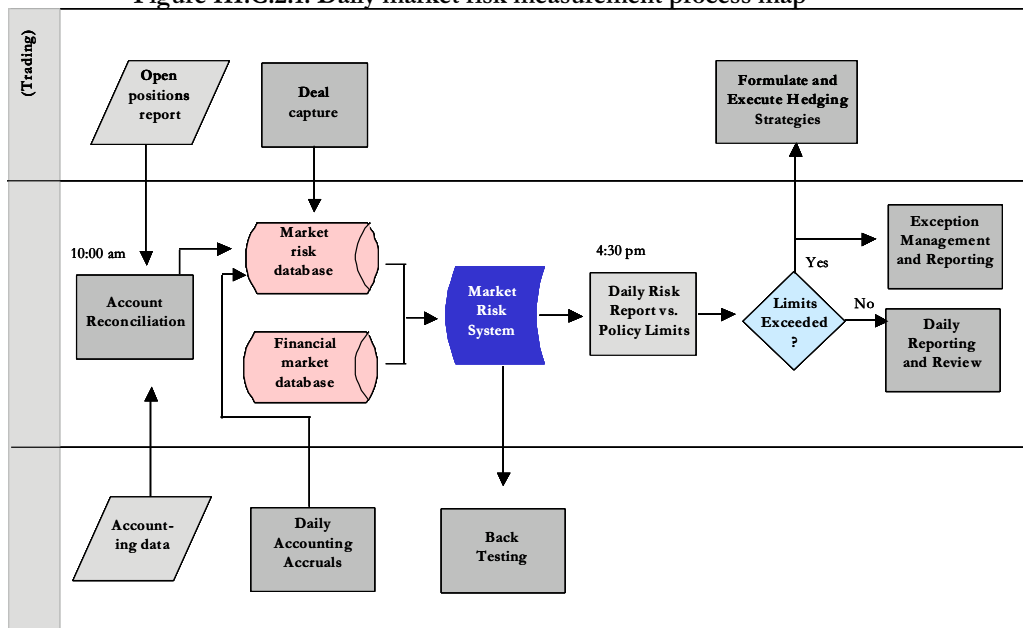
In developing and applying operational risk process models, risk managers should first take advantage of other related programmes that can provide valuable information or tools. While the discipline of ORM is relatively new, businesses have always had to ensure that their operations are effective and efficient. As such, many companies have implemented programmes to identify, monitor, and improve their business processes. These programmes often fall under the monikers of re-engineering or total quality management, and these corporate-wide efforts produce detailed process maps and performance metrics. In addition to process improvement, companies have also implemented risk assessment processes to identify key operational risks. These risk assessments are either performed by the business and operating units themselves (known as control self-assessments), or by independent internal or external audit groups.

As a starting point, the methodologies and results from these initiatives – process maps, performance metrics, audit ratings – can be used to gain a deeper understanding of the general scope and specific issues that the ORM program must address. With this knowledge, the development of operational risk process models should include the following four steps:

- *Step 1: Establish the objectives and requirements of key stakeholders.* The design of operational risk process models should always start with the end goal(s) in mind: what are the key business and operational objectives for the company? These objectives can generally be grouped into three categories, business performance, financial performance, and compliance. Business performance objectives include product innovation, customer acquisition and retention, and market share. Financial objectives include earnings growth, risk-adjusted profitability, and shareholder value. Compliance objectives should encompass internal risk policies and limits, as well as external regulatory and legal requirements. For example, one of the key objectives of a capital markets trading business is to maintain their market risk exposures within board-approved risk policy limits.
- *Step 2: Identify the core processes that support these objectives.* Most companies view their businesses *vertically* in terms of operating units, support functions, products, or customer segments. However, companies must manage their business processes *horizontally* to fully address the operational risks that may prevent them from achieving their key objectives. This is because the core processes of any company – customer acquisition, product delivery, cash

management, etc. – involve the participation of various entities within and outside of the organization. To better understand these linkages, process maps should be developed for the core processes of the company. These process maps should be driven by the objectives of the company, and highlight the specific interdependencies, such as work flows, data flows, and/or cash flows. It is also important to note that risk management is a process. As an example, Figure III.C.2.1 shows a process map for daily market risk measurement. This process map shows the work flows between the front office (traders), the middle office (risk management), and the back office (accounting and IT). It also shows two time-critical objectives: an account reconciliation between the open positions report from the front office and the accounting report by 10 a.m., and a daily market risk report showing risk exposures against limits by 4.30 p.m.

Figure III.C.2.1: Daily market risk measurement process map



Step 3: Define performance and risk metrics, including goals and MAPs. For each core process of the company, performance metrics and risk metrics should be clearly defined. For example, the systems availability of a core application is essential for day-to-day operations. A company might set 100% systems availability as a goal and 99.99% as minimum acceptable performance (MAP). Similarly, goals and MAPs should be established for all key performance and risk metrics. As such, all of the company's operations can be monitored against specific benchmarks. Over time, management can respond proactively to specific processes that perform below MAP. For processes that perform consistently above goal, then the goal and MAP for those processes can be raised to encourage continuous improvement. To follow on with our example, an operational risk metric for the daily market risk measurement process may be the percentage of time that the daily market risk report is produced by 4.30 p.m. Management can establish 99% as the goal and 95% as the MAP. As such, the goal is to produce the daily market risk report by 4.30 p.m. on 250 out of the 253 trading days in a year, with a MAP of 240 days.

Step 4: Implement organizational and risk mitigation strategies. With a clear understanding of stakeholder objectives and supporting core processes, and performance of those processes against performance standards, the company is well positioned to execute the appropriate ORM strategies. These strategies may include: new training programs; new IT applications; process redesigns; management restructuring; integration of audit, compliance, security and ORM activities; specific investigations and corrective actions; and risk transfer through insurance programmes.

In our example, suppose management noticed that the daily market risk report was late 4 times in a month, a below-MAP performance given that the frequency is greater than 13 times per year. An investigation revealed that the main reason, or root cause, is that the traders are late in updating their daily trades. Risk mitigation strategies may include discussion forums to resolve any misunderstandings or conflicts, or hiring new trading assistants to support the traders. General Electric is well known for its 'workouts' in which cross-functional teams are organized to discuss and resolve any operational issues in an open forum. To highlight the importance of this process, senior executives usually attend the last session of these workouts, to obtain an in-person report from the team leaders on how they plan to address any outstanding issues.

More sophisticated companies go beyond these four steps in their ORM programmes, and allocate economic capital to each business unit based on its operational risks. The direct linkage between capital requirements and operational risks is one of the key developments in Basel II. The allocation of capital to operational risks provides a number of benefits:

- Management can measure risk-adjusted profitability consistently across different business units and products. In fact, performance models that do not fully adjust for risks (such as economic value added models) would overstate the profitability of high-risk businesses and understate the profitability of low-risk businesses.
- Organizational incentives, in the form of lower capital charges, are provided to business units that effectively manage their operational risks. One of the key objectives of any risk model is to motivate appropriate behaviour.
- In the evaluation of risk transfer strategies, such as insurance, management can compare the cost of risk retention (i.e., economic capital times the cost of capital) and the cost of risk transfer (i.e., net cost of the insurance strategy).

As we will discuss later, the allocation of economic capital to operational risk will enhance the evaluation of the costs and benefits of these strategic alternatives.

III.C.2.3 Specific Tools

Given the wide scope of operational risk, a company should employ a range of qualitative and quantitative tools to assess, measure, and manage operational risks. Below is a summary of the basic ORM tools that companies use today:

(i) Loss-incident database. A company should record operational losses and also keep a record of operational incidents for two main reasons. First, losses are measurable and can be used to indicate trends (e.g., trend in the loss/revenue ratio). Incidents record other events that should be noted, even if they did not result in an operational loss. Second, every loss and incident within a company represents a learning opportunity, without which past mistakes are more likely to be repeated. As such, the loss-incident database should be used to support the identification of operational risk exposures, the development of risk-based audits (in which high-risk business units are audited more frequently), as well as to facilitate the sharing of lessons learned within the company. Additionally, there are several industry initiatives to develop more robust loss-event databases, but it is too early to tell which one(s) will become the industry standard. It is unlikely, however, that the management of operational risk will ever become a wholly data-driven process; given the nature of operational risk, it will always be more of a management issue than a measurement issue.

(ii) Control self-assessment. A control self-assessment (as distinct from a risk self-assessment – see Section II.C.2) is an internal, subjective analysis of the key risks, the controls available to mitigate these risks, and the management implications. It is important for all of the business units to assess their current situation in terms of a control self-assessment to develop a clear picture of

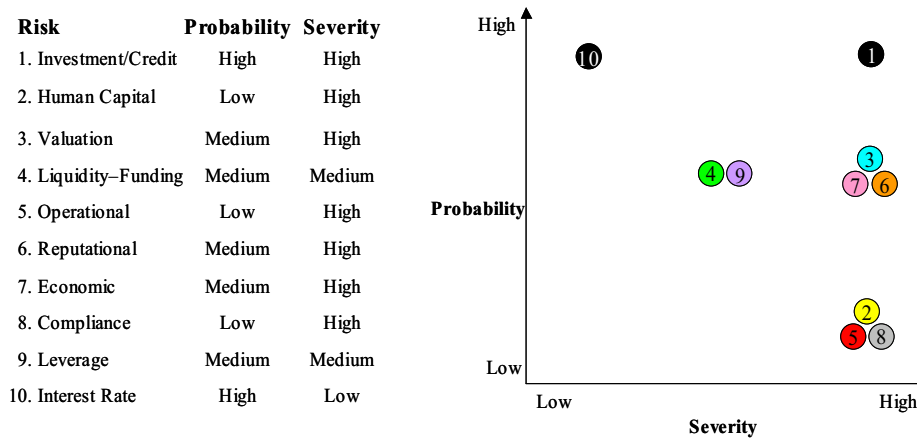
how to proceed in the ORM process. Given that they fully participated in the assessment process, they would also have a greater sense of ‘ownership’ to address outstanding opportunities or issues. Tools that support self-assessments include questionnaires, issue-specific interviews, team meetings, and facilitated workshops. The output is an inventory of key risk exposures, key control initiatives, and sometimes even a Letterman-style ‘top 10 risks’. The following are questions that might be included in a control self-assessment:

1. What are the key business and financial objectives for the business unit in the next 12 months?
2. What are the key risks that may prevent you from attaining these objectives?
3. What policies, procedures and controls do you have in place to ensure that risks are within acceptable levels?
4. What new risk management initiatives do you have planned for the next 12 months?
5. What metrics, tests and reviews provide you with assurance that the policies, procedures and controls specified in 3 and 4 above are indeed effective?

(iii) Risk mapping. Building on the work from control self-assessments, the company’s key risk exposures can be ranked with respect to their ‘probability’ and ‘severity’ so that management can have a comparative view in the form of a two-dimensional risk map. Figure III.C.2.2 shows an example of a risk map. Some ORM professionals argue that companies should be most concerned about events of low probability and high severity because management does not have sufficient experience in dealing with these events, whereas they have more experience in dealing with events of high probability and high severity. For operations that are more complex (e.g., outsourcing arrangements, special-purpose vehicles), risk-based process maps can be produced to show how various risk exposures can arise. These maps will aid in the identification of the risks encountered in each business unit, indicating ‘problem spots’, such as single points of failures or where errors often occur. These maps will also enable each business unit to develop and prioritize their risk management initiatives to address the most important risks.

(iv) Key risk indicators. Risk indicators are quantitative measures that are linked to operational risks for a specific process. Examples include customer complaints for a sales or service unit, trading errors for a trading function, unreconciled items for an accounting function, or system downtime for an IT function. These risk indicators are usually developed by the individual business units and closely tied to their business objectives. Early-warning indicators should also be developed to provide management with leading signals (e.g., employee absenteeism and turnover as an early warning indicator of future operational errors). As discussed earlier, the establishment of goals and MAPs will provide useful performance benchmarks against which the key risk indicators can be measured.

Figure III.C.2.2: Risk Map



Other sources of valuable information for risk identification and assessment include internal audit reports, external assessments (external auditors, regulators), employee exit interviews, and customer and employee surveys.

Operational risk professionals also find it useful to distinguish between key risk indicators (KRIs) and key risk drivers (KRDs). KRIs are ex-post indicators of operational risk performance, in that management has no direct control over their outcomes. On the other hand, KRDs are levers that management has direct control over. Examples of KRDs include number of training hours, number of automated versus manual processes, time to fill open positions, and time to resolve outstanding audit findings. As such, KRDs can be best thought of as controllable factors that will influence future KRIs.

III.C.2.4 Advanced Models

When ORM first came on the scene a few years ago, there were basically two distinct schools of thought. One school subscribed to the notion that you cannot manage what you cannot measure, and they focused on quantitative tools such as loss distributions, risk indicators, and economic capital models. The other school believed that operational risk cannot be quantified effectively, and they focused on more humanistic, qualitative approaches such as self-assessments, risk maps, and audit findings.² Today, operational risk professionals realize that best practices must integrate both quantitative and qualitative tools.

² See Lam (2003a).

In addition to the basic risk identification and assessment tools discussed above, leading companies employ advanced operational risk models. Unlike market risk and credit risk, where risk measurement methodologies have been developed and tested for many years, there are no widely accepted models for operational risk measurement. In selecting a methodology (or combination of methodologies), each company should first establish its objectives and resources and choose accordingly. Different methodologies imply different interpretations of operational risk, and require various inputs to be useful. Given that there is likely to be no single solution, a combination of methodologies will allow the disadvantages of one model to be balanced by the strengths of another, allowing a more robust overall measurement to be developed. Some of the most common methodologies, including their strengths and weaknesses, are discussed in this section (see Hernandez *et al.*, 2000, for a detailed discussion of key strengths and weaknesses of various ORM models).

III.C.2.4.1 Top-down models

The top-down approach to operational risk assessment calculates the ‘implied operational risk’ of a business by using data that are usually readily available, such as the overall financial performance of the company or that of the industry in which it operates. Top-down models use relatively simple calculations and analyses to arrive at a general picture of the operational risks encountered by a company. These top-down models benefit from the sophisticated methodologies already developed for credit and market risk. Examples of top-down models on operational risk include the implied capital model, the income volatility model, the economic pricing model, and the analogue model:

(i) Implied capital model. This methodology assumes that the domain of operational risk is ‘that which lies outside of credit and market risk’. Thus, the capital allocated to operational risk must be the result of subtracting the capital attributable to credit and market risk from the total allocation of capital. Although this model provides an easily calculated ‘number’ for operational risk, its simplicity presents several disadvantages. First, total risk capital must be estimated given the company’s actual capital and the relationship between its actual debt rating and target debt rating. Second, it ignores the interrelationships between operational risk capital and market risk and credit risk capital. Finally, this model does not explicitly capture the causes and effects for operational risk.

(ii) Income volatility model. This model is similar to the capital allocation model, but it goes one step further, by looking at the primary determinant of capital allocation – income volatility. The volatility attributable to operational risk is calculated in the same way as in the capital allocation model – by subtracting the credit and market risk components from the total income volatility.

One of the advantages of this model is that of data availability: historical credit and market risk data are usually easily obtained, and total income volatility can be observed. However, this model also has several shortcomings, the most dramatic of which is that it ignores the rapid evolution of firms and industries. Structural changes, such as new technologies or new regulations, are not captured in this model. The income volatility model also fails to capture softer measures such as opportunity costs or reputation damage. In addition, it fails to capture the low-frequency, high-severity risks, as is true in all of the top-down approaches.

(iii) Economic pricing model. The capital asset pricing model (CAPM) is probably the most widely used of economic models, and can be used to determine a distribution of the pricing of operational risk relative to the other determinants for capital (see Chapter I.A.4). The CAPM assumes that all market information is captured in the share price; thus the effect of publicized operational losses can be determined by evaluating the market capitalization of a company. The advantage of this approach is that it incorporates both discrete risks and softer issues such as reputational damage and effects of forgone opportunities. With this approach, a company's stock price volatility due to operational risk is derived by taking the company's total stock price volatility and subtracting from it the stock price volatility due to credit risk and market risk. However, the CAPM approach presents an incomplete and simplistic view of operational risk. It provides only an aggregate view of capital adequacy, not information about specific operational risks. Furthermore, the level of operational risk exposure is not affected by particular controls and business risk characteristics, so there is no motivation to improve operations, and while tail-end risks *are* incorporated in the model, they are not thoroughly accounted for. This is a significant omission. Such incidents can do more than just diminish the value of a business: they can lead to the end of the business completely. Finally, this model does not help in anticipating, and therefore avoiding, incidents of operational risk.

(iv) Analogue model. The analogue model is based on the assumption that one can look at external institutions with similar business structures and operations to derive operational risk measures for one's own organization. This model can be extended to look for the causes and effects of operational losses at such institutions. This method offers one way to proceed when a company does not have a robust database of operational risk losses. However, it takes some credulity to assume that the high-level numbers of another institution can accurately measure one's own operational risk, and many are suspicious of this approach. In the words of one analyst: '[The] intangibles within an institution – its risk-taking appetite, the character of its senior executives, the bonus structure of its traders – put so many wild cards into the operational risk equation that similarities in business volume, transaction volume, documented risk policies and other qualities that can be scored are swamped.'

III.C.2.4.2 Bottom-up models

The bottom-up methodology applies loss amounts and/or causal factors to predict operational losses in the future. It requires a company to clearly define the different categories of operational risk that it faces, gather detailed data on each of these risk categories and then quantify the risk. A company often needs to augment its internal data with an external loss-event database. The final output of this bottom-up approach is a loss distribution that enables operational risk capital to be estimated for a given confidence level (see Chapter III.C.3). A number of surveys have indicated an increasing preference for risk-based bottom-up methodologies over the top-down approaches. The Basel II requirements should further encourage banks to develop bottom-up models. The data needed for this methodology can also be used to derive a business risk profile. For example, turnover or error rates can be tracked over time and combined with changes in business activities to construct a more robust picture of the business operational risk profile. By tracking these KRIs over time, the company can assess its operational risk exposure on an ongoing basis and can upgrade specific controls as needed. Furthermore, continuous tracking provides a company's management with better information about its operations and increases awareness of the causes of operational risk.

However, bottom-up models present several difficulties. Mapping loss data from the company with loss data from other companies is complex, given the differences in business mix, size, scope and operating environment. Even mapping internal losses to specific risk types is difficult because losses are frequently reported as aggregates from multiple risk sources that are difficult to isolate. For example, an operational loss on a trading floor might result from personnel risk, lack of trading controls, expanding overseas business, lack of back- and front-office segregation, volatile markets, senior management confusion, and incompetence. In addition, robust internal historical loss data may not be available, particularly for low-frequency, high-severity events.

Bottom-up models are usually based on statistical analysis and scenario analysis. Classical statistical models require an ample supply of operational loss data that are relevant to the business unit. The lack of appropriate internal data is therefore the greatest obstacle to the widespread application of this methodology; the use of external data as a proxy poses several problems, as mentioned earlier. However, the analytical power of this tool will hopefully become more widely applicable in the near future as increased awareness of operational risk leads to improvements in data collection and extensions of the classical statistical methodology.

Scenario analysis offers several benefits that are not addressed by the classical statistical models. A scenario analysis is used to capture diverse opinions, concerns and experience/expertise of key managers and represents them in a business model. Scenario analysis is a useful tool in capturing

the qualitative and quantitative dimensions of operational risk. Risk maps allow the representation of a wide variety of loss situations, and capture the details of the loss scenarios envisioned by the managers surveyed. Risk maps of each business unit identify where operational risk exposures exist, the severity of the associated risks, whether any controls are in place, and the type of control: damage, preventive, or detective. Cause and effect relationships can be captured with this methodology. The shortcoming of such a model, however, is in its subjectivity, which creates a potential for recording data inconsistently and/or for biasing conclusions if one is not careful.

At the beginning of this section we discussed the need to balance the qualitative and quantitative tools. For example, control self-assessments require that business units are honest and forthright about their major operational risks, which can often be embarrassing problems that they would rather not discuss (let alone highlight for senior management!). To counterbalance this shortcoming, business units should be required to not only ‘tell me’ but also to ‘show me’. This can be accomplished through validation processes, such as:

- pre-established operational risk indicators that are monitored against goals and MAPs;
- periodic tests to ensure that actual losses and incidents (ex post) result from operational risks that were being monitored through KRIs or at least discussed in the control self-assessments (ex ante);
- comparisons between control self-assessments and independent assessments such as internal audits, external audits, regulatory reviews, and customer surveys.

In fact, Section 404 of the Sarbanes-Oxley Act requiring management assessment of internal controls for financial reporting, as well as auditor attestation, was designed to ensure such validation.

III.C.2.5 Key Attributes of the ORM Framework

Today, ORM practitioners recognize the pitfalls of using only one approach to modelling operational risk – either top-down or bottom-up – and that best practice ORM incorporates elements of both approaches. We will now discuss the attributes of a unified ORM framework, and then how these attributes can underpin a seven-factor economic capital model.

A unified ORM framework should satisfy two basic requirements. First, it should support both the measurement and management of operational risks. Second, the ORM framework should incorporate the interdependencies across credit, market and operational risks as part of an overall ERM program. Based on these two requirements, the key attributes of a unified ORM framework include the following:

(i) Integrating qualitative and quantitative tools. The nature of operational risk (i.e., the risk of loss due to people, processes, systems and external events³ is complex and dynamic. As such, the advantage of qualitative tools is that they can incorporate human experience and judgement in order to capture risks that are subjective. For example, what are the operational risks associated with a new product? On the other hand, the advantage of quantitative tools is that they provide objective indicators that can be used to show aggregate losses, exposures, and trends against established targets. A unified ORM framework should incorporate both advantages, as well as integrate the institution's various risk management and oversight activities (e.g., ORM, audit, compliance, quality, insurance).

(ii) Providing early warnings and escalations. Operational risk cannot be managed effectively based only on backward-looking indicators such as losses, error rates, and incidents. The ORM framework should provide early warning indicators of emerging risk issues. A quantitative example is that an increase in employee absenteeism may be an early warning for increasing turnover and human errors. A qualitative example is competitive intelligence that indicates significant investments in a new technology by a key competitor that, if successful, would render the firm's existing technology obsolete. An ORM framework should establish early warning indicators, as well as effective escalation processes so that management can take the appropriate actions. For example, a money management company established specific escalation processes such that the higher the number of customers impacted by an incident, the higher the level of management is notified. This ensures that 'bad news travels up' the organization and that the appropriate level of management responds in a timely manner.

(iii) Influencing business activities. One of the most important attributes of an ORM framework is that it influences business actions and decisions. Such influence can be asserted through: (1) corporate policies with respect to guidelines for, and restrictions on, business activities, such as acceptable versus unacceptable sales practices; (2) teamwork between the line units and ORM in new business and product development processes; (3) risk response plans based on ORM indicators and escalations; (4) adjustments in economic capital given operational risk performance and risk mitigation strategies; and (5) positive and negative incentives to motivate appropriate business behaviour. This attribute ensures that operational risks are managed on an ongoing basis, and that specific consequences are in place to provide organizational reinforcements. An excellent example of using positive incentives is when GE tied one-third of senior management compensation to the achievement of quality management objectives as part of the company's 'six sigma' programme.

³ The definition of operational risk in this chapter includes business risk, which is notably absent in Pillar I of the Basel II proposals.

(iv) Reflecting environmental changes. Just as credit risk and market risk frameworks reflect changes in underlying default rates and market prices, an ORM framework should reflect changes in the operational risk environment. For example, increases in industry-wide operational risk losses and incidents may indicate an increase in systemic risk. A number of industry loss-event databases are being developed that can provide this type of information. Other environmental changes include new legal and regulatory requirements, such as those established by the Sarbanes-Oxley Act, the Patriot Act and the Basle II proposals. A company that lacks the processes and systems to comply with these new requirements is likely to face greater operational risk with respect to regulatory scrutiny and legal penalties.

(v) Incorporating risk interdependencies. There are important interdependencies within and across risk types. For example, credit risk is the primary concern for most banks, but inadequate loan documentation (an operational risk) is likely to increase loss severity in the event of a borrower default. An ERM programme should address such interdependencies in the design of early warning indicators, the development of scenario analysis, and the implementation of risk response plans. For example, financial institutions must simultaneously manage market risk and operational risk during stressed market conditions (e.g., the Russian crisis during the autumn of 1998). Given that there is a high correlation between volatile prices (a key driver for market risk) and transactional volumes (a key driver for operational risk) during stressed periods, financial institutions should establish early warning indicators and risk response plans. Examples of early warning indicators are shown in Figure III.C.2.3. If these indicators exceed a critical level, management should implement pre-established contingency plans, such as reduction of trading limits to reduce market risk exposures and activation of back-up sites to increase processing capacity.

Figure III.C.2.3: Early warning indicators

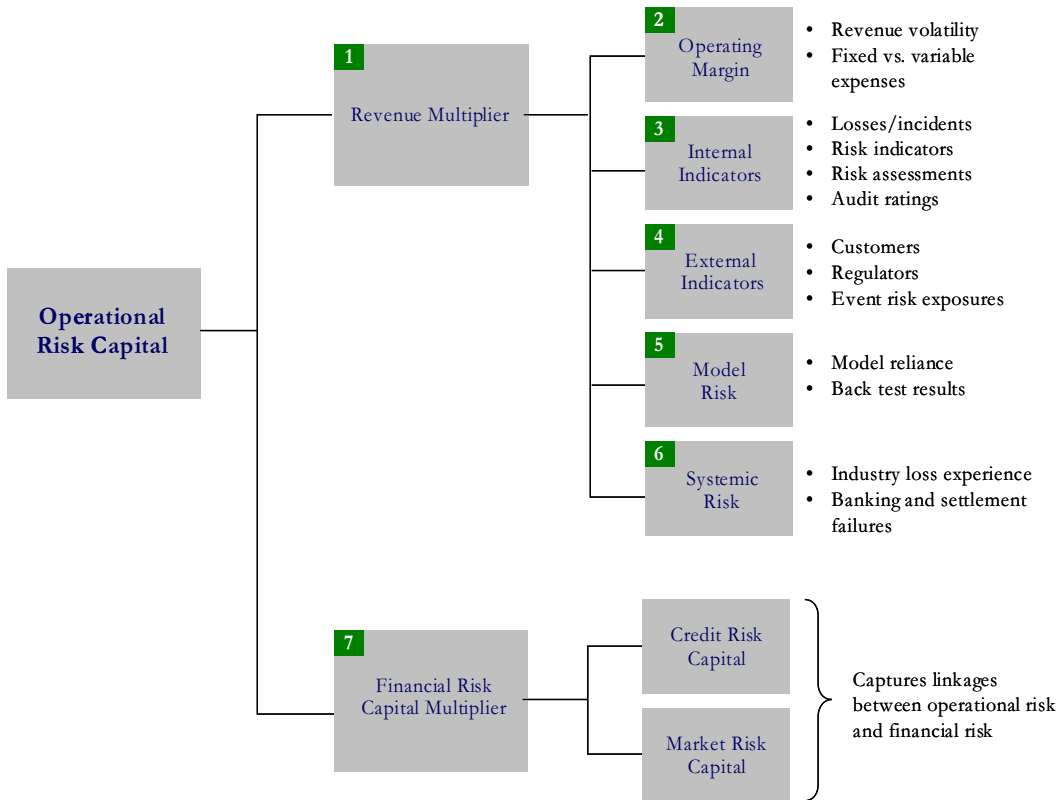
Risk Category	Early warning indicators
Credit Risk	<ul style="list-style-type: none"> • Borrower/counter party stock price declines • Widening of credit spreads in the debt and credit derivatives markets
Market risk	<ul style="list-style-type: none"> • Increases in actual and implied price volatilities • Breakdowns in historical price relationships and patterns
Business/Operational Risk	<ul style="list-style-type: none"> • Spikes in business growth, profitability, and complexity/change • High and undesirable turnover rates
Enterprise-wide Risk	<ul style="list-style-type: none"> • Increases in any risk concentrations and/or organizational powers • Changes in intra- and inter-risk correlations

As we will discuss in the next section, these interdependencies should also affect the determination of economic capital, and in ways that might not be obvious.

III.C.2.6 Integrated Economic Capital Model

Given the five attributes of a unified ORM framework discussed above, what factors should determine economic capital for operational risk? Figure III.C.2.4 shows a seven-factor approach to calculating operational risk capital. While this is conceptual, it can be adapted to a firm's specific business mix, size, and risk profile.

Figure III.C.2.4: Operational risk capital calculation



Let us discuss each one of these factors in turn.

(i) *Revenue multiplier.* This is a top-down estimate of the amount of operational risk capital required by a business or operating unit. Such an estimate can be derived from observing analogues of publicly traded companies in same or similar businesses, while adjusting for market risk and credit risk. For example, Capital One may be a credit card company analogue, while First Nationwide may be one for mortgage companies. Outsourcing firms such as IBM or EDS may be analogues for internal IT functions. The central question is ‘if the business or operating unit were a standalone business, how much capital would it need for operational risk capital?’ The revenue multiplier⁴ assumes an average operational risk profile, which can then be adjusted upwards or downwards by the other factors below.

(ii) *Operating margin.* This factor incorporates the degree to which the firm’s operating margin is more or less volatile than average, and is often referred to as ‘business risk’. A firm’s inability to

⁴ For certain businesses, a top-down proxy based on activity or volume might be more appropriate.

generate sufficient revenue to cover expenses (net of unexpected credit and market risk losses) is a major reason why it needs to hold operational risk capital. For example, business variables that can increase the required operational risk capital include greater volatility in business volume, weak power to set prices, and higher fixed versus variable expenses.

(iii) Internal indicators. This adjustment reflects the effectiveness of internal controls. A scorecard should be developed for the internal quantitative and qualitative indicators, with individual weightings, that would provide an overall adjustment to operational risk capital. Internal indicators would include losses, incidents, risk metrics (e.g., error rates, unreconciled items), early warnings, internal audit ratings, risk maps, etc. The economic impact of contingency plans and insurance programmes should also be factored in. Each key indicator should also be associated with specific goals and MAPs.

(iv) External indicators. As with internal indicators, a scorecard of external indicators should be developed. External indicators would include customer satisfaction scores and complaints, external audit comments, and regulatory exam findings. This scorecard would also track exposures to external events, such as fires, earthquakes and acts of terrorism. Firms that rely on external vendors should also incorporate vendor performance relative to service level agreements. Goals and MAPs for external indicators should also be established.

(v) Model risk. This factor reflects the degree to which a firm relies on models, and the quality of such models. The primary input is back-testing results against predetermined criteria. A firm should include all models that drive management decisions and actions, such as pricing and valuation models, scenario and simulation models, and risk management models. For firms that do not rely on models, this may simply be one of the internal indicators.

(vi) Systemic risk. This factor adjusts for dramatic shocks in the business environment, such as industry-wide losses and incidents, and banking and settlement failures. Systemic risk is especially important for highly interconnected industries such as financial services and energy services, where trading activities and counterparty exposures within the industry are significant. Past examples include the Long-Term Capital Management collapse, Y2K readiness, and the Enron bankruptcy. In each of these situations, companies were concerned not only about their direct exposures, but also the exposures of their business partners and counterparties.

(vii) Financial risk multiplier. This factor is meant to capture the compounding effects between operational, credit, and market risks. It is not portfolio diversification, which may lead to a reduction in aggregate economic capital at the enterprise-wide level. In fact, it is a compounding

factor that many risk managers ignore. Regulators refer to this compounding factor as ‘spillover effects.’ Cumming and Hirtle (2001) argued that the confluence of variables including market liquidity problems, lack of corporate limberness, and reputational and contagion effects, could result in the aggregate risk of a firm *exceeding* the sum of its individual risks. The financial risk multiplier is meant to capture such spillover effects. An argument can also be made that a variety of operational risk exposures (e.g., rogue trader, inadequate loan documentation, unsavoury sales practices) are compounded in a firm with significant market risk and credit risk exposures. After all, a rogue trader can do much more damage at a bank than at a retail store.

The practice of ORM has come a long way in the past several years. It still has a long way to go. At the annual 2003 operational risk conference organized by the Risk Management Association, Eric Rosengren of the Federal Reserve Bank of Boston said that only three of the 20 largest US banks qualify for the ‘advanced management approach’ for operational risk under Basle II, which is supposed to lead to reduced capital charges. However, the development of ORM is more than a regulatory compliance issue. Early adopters of more sophisticated ORM have reported significant business benefits, including improved customer service, greater operating efficiency and reduced losses. To fully realize these benefits, it is clear that the further development of ORM practices must integrate quantitative and qualitative tools.

III.C.2.7 Management Actions

Assessing and measuring operational risk is important, but pointless unless directed towards the improved management of operational risk by enhancing internal controls and controlling key risk factors. Simply stated, the goal of ORM is to help management to achieve their business objectives. Once a measurement framework is in place, the next step is to implement a process that identifies actions that will reduce operational losses. These actions include adding human resources, increasing training and development, improving and/or automating processes, changing organizational structure and incentives, adding internal controls (e.g., more frequent or more extensive monitoring), and upgrading systems capabilities.

The key to effective operational risk mitigation is to establish a cross-functional rapid response team that will address and resolve any emerging operational risk issues. At one business unit at Fidelity Investments these teams were called ‘turbo teams’, and they responded immediately when operational risk indicators fall below MAP and reported back to management on their assessments and actions within a few days or weeks.

Finally, a mechanism for evaluating and prioritizing potential improvements must be created. Cost–benefit analysis and readiness assessments are useful tools that should be included in the

evaluation process. For example, business executives often turn to IT, or more specifically automation, as the answer to process improvements. However, the potential benefits must be weighted against the total costs of the project (e.g., development, testing, training, implementation and ongoing maintenance). More importantly, management must ensure that the organization is ready to take advantage of the technology solutions. Automation of poorly designed processes can result in significant operational risks in the future.

Some of the operational risk measurement approaches discussed above should naturally lead to improved operational risk management at the business unit level. A business unit can monitor and improve its operational risk levels by setting operational goals, exposure limits and MAPs on key operational processes. For example, suppose a brokerage group on average processes a million trades per day. This group may specify that its operational goal is that failed trades be less than less than 50 per day, while its MAP is no more than 100 failed trades per day. Additionally, the group may specify that no more than 40% of daily trades can be processed by one operational centre in order to spread its reliance across multiple operational centres.

The allocation of economic capital for operational risk, if it captures both performance and behaviour effects, should motivate business units to improve their ORM in order to reduce their capital charges. For example, a business may set up procedures through which employees may respond immediately to operational problems and implement the controls necessary to monitor and improve performance. A key requirement for risk mitigation is to understand the root causes of operational risks, such as lack of training or inadequate systems, and then focus corrective actions on these root causes. Business units that take the appropriate actions should receive a 'credit' in their economic capital charges.

One of the key objectives of any ORM programme is to ensure that 'bad news travels up an organization'. In other words, the quantification and modelling of operational risk should lead to more timely communication and escalation of operational risk issues. This can be accomplished through the various process models and quantitative tools discussed above. Additionally, management should clearly communicate when they should be informed through a cascading set of 'escalation triggers', which would lead to the appropriate decisions and actions on the part of management. Escalation triggers can be defined in terms of the KRIs, such as the level of operational losses, the number of errors, significant policy violations, and number of customers impacted by an incident. These escalation triggers and procedures should be incorporated into the company's policies and procedures, as well as training programmes and on-line support tools, so that all employees understand what is expected of them.

Besides risk mitigation through operational processes and controls, and economic capital incentives for ORM, there are other financial solutions that management may consider. Companies can establish reserves to cover their expected operational losses. These reserves are considered a form of self-insurance. Expected losses should be embedded in the pricing of a product, indeed market and credit risks are already incorporated into some transaction prices as a matter of practice. Including an additional adjustment for operational risk makes for a more comprehensive picture and allows for more accurate risk-adjusted pricing. For example, if a business unit performs 10,000 transactions annually, with an expected loss of \$80,000 a year due to operational factors, then an adjustment of \$8 per transaction could cover such losses.

Additionally, the cost of capital for operational risk (and other risks) should be incorporated into the pricing of a transaction. Pricing can also be driven by the target levels of returns that the company expects a product to achieve given competitive pricing and market share objectives.

III.C.2.8 Risk Transfer

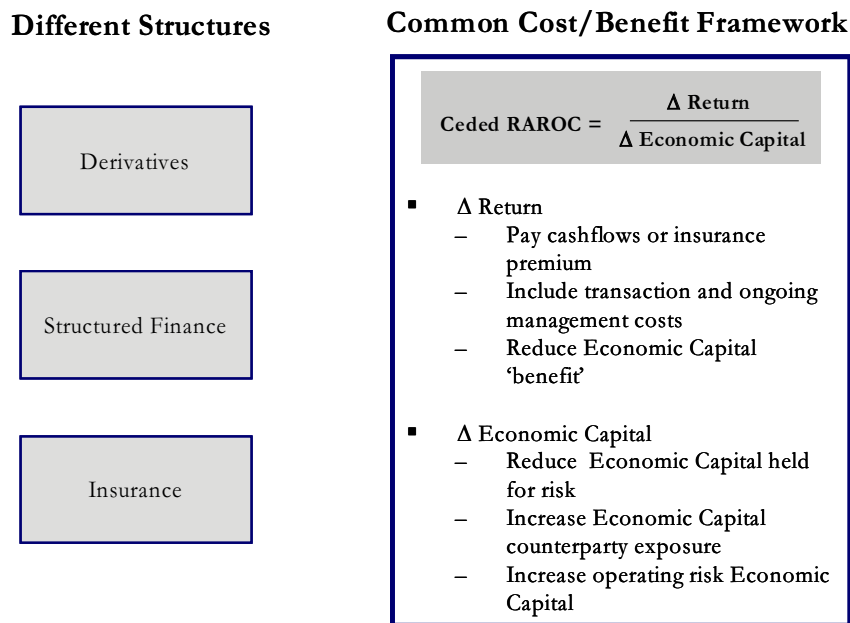
For critical operational risk exposure, a company must decide if the best strategy is to implement internal controls and/or execute risk transfer strategies. The two are not mutually exclusive and are often complementary. For example, most companies implement workplace safety procedures (an internal control) and purchase workers' compensation insurance (a risk transfer strategy). In fact, the former can reduce the cost of the latter. Another example is product liability. A company can strengthen product development controls as well as purchase product liability insurance.

Some risk transfer strategies are intended as 'backstops' to internal controls. For example, directors' and officers' liability insurance provides protection against 'wrongful acts'. In the past, insurance managers would purchase such 'backstop' insurance policies based on the structure, cost, and provider rating and service level. In the context of ERM and ORM, a company should:

- identify its operational risk exposures and quantify its probabilities, severities and economic capital requirements;
- integrate its operational risk with its credit risk and market risk in order to assess its enterprise-wide risk–return profile;
- establish operational risk limits (e.g., MAPs, economic capital concentration);
- implement internal controls and develop risk transfer and financing strategies;
- evaluate alternative providers and structures based on cost–benefit economics (i.e., comparing the cost of risk retention and risk transfer).

The economic capital framework discussed above is also a useful tool for evaluating the impact of different risk transfer strategies. For example, in executing any risk transfer strategy the economic benefits include lower expected losses and reduced loss volatility, while the economic costs include insurance premiums, as well as higher counterparty credit exposures. In a sense, the company is both ceding risk and ceding return, resulting in a ‘ceded RAROC’. By comparing the ceded RAROCs of various risk transfer strategies, a company can compare different structures, prices and counterparties on an apples-to-apples basis and select the most optimal transaction(s). Moreover, a risk transfer strategy with a ceded RAROC below the firm’s cost of equity would add to shareholder value, and vice versa. Figure III.C.2.5 provides the framework for a ceded RAROC analysis.

Figure III.C.2.5: Ceded RAROC analysis



Let us apply the framework using the following example of purchasing a product liability insurance policy at an annual premium of \$100,000:

1. Annual insurance premium	\$100,000
2. Annual management cost	20,000
3. Economic capital benefit ⁵	60,000

⁵ The economic capital benefit represents a funding credit. It is used in matched-maturity funds transfer systems to recognize the interest income from the investment of capital funds. In our example, we assume a funding rate of 3%.

4. Net reduction in economic capital ⁶	2,000,000
---	-----------

In this example, the ceded RAROC is 9% $[(100,000 + 20,000 + 60,000) / 2,000,000]$. This represents the effective cost of risk transfer, which can be compared to the effective costs of alternative risk transfer strategies as well as the cost of risk retention. For instance, if the company's cost of economic capital were 10%, then this transaction would add to shareholder value because the ceded RAROC (cost of risk transfer) is below the cost of risk retention.

III.C.2.9 IT Outsourcing

IT outsourcing is widely considered one of the major business imperatives in today's business world, often described as a 'mega-trend'. The META Group estimates that IT outsourcing is a US\$150 billion market, compared to \$200–220 billion for software and hardware. More importantly, the growth of IT outsourcing is expected to outpace that of software and hardware for the next few years at least. Let us discuss how one might establish an operational risk process for IT outsourcing based on the processes and tools discussed earlier.

III.C.2.9.1 Stakeholder Objectives

Buyers of outsourcing should first establish an overall outsourcing strategy. This strategy would identify the IT systems and/or applications that are outsource candidates, the approach to evaluating alternative outsource providers and solutions, and the decision processes and cost-benefit analyses that will result in specific outsourcing transactions. This strategy should also discuss how the outsourcing strategy will support the overall business strategy of the company, including the specific business and financial objectives that outsourcing is expected to achieve. Buyers of outsourcing services often cite the following expected benefits:

- cost savings;
- access to IT skills and advanced technologies;
- quality of services;
- resource allocation to core activities;
- scalability and flexibility in IT resources;
- shorter time to market;
- enhance e-business applications.

As a sign of more realistic expectations on the part of buyers, expected cost savings cited by buyers have come down from 40–70% a few years ago to 10–20% today. However, even with reduced expectations, outsourcing arrangements often fail to meet the buyers' requirements. A

⁶ The net reduction in economic capital includes the gross reduction of economic capital for the risk exposure that is being insured, minus the increase in counterparty and operational risk capital.

2004 IT outsourcing study by DiamondCluster⁷ noted that 21% of buyers said that they had prematurely terminated an outsourcing agreement in the last 12 months. The most common reasons cited for cancelling outsourcing arrangements include:

- provider having financial difficulties (credit risk);
- provider failure to deliver on commitments (operational risk);
- buyer consolidation of outsourcing vendors (business risk).

In addition to these risks, buyers must address a complex set of significant operational risks – such as geopolitical risk, regulatory compliance, data integrity and security, and reputational risk. While project delays and cost overruns are the most common reported outcomes from these risks, it might be useful to review a couple of more specific examples. In the area of information security, a woman from Pakistan recently obtained sensitive patient information from the University of California, San Francisco, Medical Center through a medical transcription subcontractor that she worked for, and threatened to post the files on the Internet unless she was paid more money. On the political side, there is a significant backlash against ‘exporting jobs’ through outsourcing. In response to political pressure, the state of New Jersey has passed legislation to ban IT and business process outsourcing by state government.

Perhaps one of the most critical risks in IT outsourcing arrangements is the appropriate alignment of objectives between the parties. Given the long-term nature of outsourcing contracts, buyers must seek out a provider that not only offers the optimal technologies and services, but also possesses compatible business culture, professional standards, and business objectives.

III.C.2.9.2 Key Processes

The key processes associated with IT outsourcing include:

(i) Evaluation and selection of outsourcing provider. The process includes documentation of requirements in a request for proposal (RFP), sending out RFPs and evaluating the responses, and negotiating the contract and service level agreement (SLA). This step can take six months to a year, and cost 2–5% of the annual cost of the contract.

(ii) Transitioning to the outsourcing environment. The transition period is perhaps the most challenging stage of an outsourcing initiative. Steps include bringing the provider company professionals on-site for training and knowledge transfer, transferring or terminating existing employees, and

⁷ DiamondCluster 2004 Global IT Outsourcing Study, available at www.diamondcluster.com.

establishing the required infrastructure (hardware, software, communication protocols) and operational processes. This step can take an additional three to twelve months, and cost 5–15% of the annual contract.

(iii) Ongoing management of the outsourcing contract. On a day-to-day basis, the buyer must manage work processes and communications, monitor provider performance, audit operations, perform quality tests, and integrate the output with other in-house or outsourced IT operations. This requires significant project management resources, and can cost 5–10% of the contract. Additionally, the ongoing ‘costs of risk’ should be considered, such as incremental insurance expense and economic capital costs.

Each of the above processes, especially ongoing management, should be fully documented in policies and procedures and illustrated in process maps. As such, roles and responsibilities of both parties are clearly established.

III.C.2.9.3 Performance Monitoring

As discussed earlier, performance and risk metrics should be developed as part of an ongoing outsourcing review process. Performance goals and MAPs should be incorporated into SLAs and a ‘scorecard’ should be developed to track actual performance against goals and MAPs. The DiamondCluster study noted that 70% of buyers and 69% of providers monitor performance at least monthly, with the following quantitative metrics as being the most common:

- on-time delivery;
- cost effectiveness;
- end-user satisfaction;
- timely, quality staffing;
- service availability;
- time to process requests;
- defect rates;
- standards compliance;
- size of request backlog.

In addition to the above performance metrics, the buyer should monitor the provider’s key business and risk metrics, such as market share, earnings, stock price performance, debt rating and financial ratios. Performance and risk monitoring should not be viewed as a reporting exercise, but as a proactive process to ensure overall project success, which may include some of the risk mitigation strategies discussed below.

III.C.2.9.4 Risk Mitigation

One of the first and most important steps in mitigating outsourcing risk is to fully evaluate the economic costs and benefits at the onset, including consideration of all critical risk factors. As noted in a survey of 500 human resources executives by Hewitt Associates, 92% of the firms had

moved jobs overseas to cut costs. However, less than half of those companies studied the tax environments of the offshore country and only 34% considered the expense of shutting down US facilities.

Let us take a simple example to see how an outsourcing contract should be evaluated. Suppose a US company is considering moving one of its application development projects offshore to China, which provides a tax-adjusted 50% labour arbitrage on a \$10 million contract. The following table shows the cost–benefit analysis for both a best case and worst case:

(\$ Thousands)	<u>Best Case</u>	<u>Worst Case</u>
1. Projected labour cost savings	\$5000 (50%)	\$5000 (50%)
2. Vendor selection	– 200 (2%)	– 500 (5%)
3. Transition costs	– 500 (5%)	– 1500 (15%)
4. Ongoing management	– 500 (5%)	– 1000 (10%)
5. Risk costs ⁸	<u>– 1000 (10%)</u>	<u>– 1500 (15%)</u>
Net savings:	\$2800 (28%)	\$500 (5%)

The above example shows that the adjusted net savings range from 28% in the best case to only 5% in the worst case. Such an analysis should be performed for different outsourcing strategies and various providers. Finally, the projected range of net savings should then be considered against less tangible risks such as reputational risks and geopolitical risks.

In addition to performing a full cost–benefit analysis of outsourcing opportunities, other risk mitigation strategies include:

- *Developing a hybrid outsourcing strategy.* Outsourcing experts suggest that the optimal blend of in-house and outsourced IT resources is generally 20–30% in-house and 70–80% outsourced. At a significant scale, even the outsourced component should be diversified across different providers, countries and locations. Other considerations include the mix between ‘nearshore’ operations (e.g., Canada, Mexico, in the case of a US firm) and offshore operations (e.g., India, China), as well as the possibility of setting up captive outsourcing companies.
- *Taking an incremental outsourcing approach.* Early outsourcing contracts were large and very ambitious arrangements that often failed to live up to the buyer and/or provider expectations. For buyers new to outsourcing or experienced buyers working with a new

⁸ Risk costs include incremental insurance expense for the outsourcing arrangement and the cost of incremental operational risk capital. The latter should include exit cost, which is a function of the probability of project failure and the cost of switching operations in-house or to another provider.

provider, a more practical approach is to outsource incrementally to ensure a compatible relationship in terms of expectations, corporate cultures, and technology and service requirements. Once the initial work is performed at a satisfactory level, then the outsourcing relationship can be expanded over time.

- *Negotiating a flexible win-win contract.* It is critical that the outsourcing contract is attractive to both parties. For example, an arrangement that is overly favourable to the buyer might not get the appropriate level of service and attention of the provider. However, the provider should have 'skin in the game' to provide the right incentives for ongoing performance. For example, the META Group estimates that by 2006, 35% of outsourcing contracts will adopt output-based pricing instead of time and materials (input-based pricing). Also, given the rapid changes in technologies, customer preferences and business requirements, contract terms should incorporate sufficient flexibility so that they do not become obsolete prematurely.
- *Establishing exit strategies and contingency plans.* Companies should establish exit strategies and contingency plans in the event that the outsourcing contract expires, or the provider does not deliver as expected, or business conditions require the termination of the contract. These exit strategies and contingency plans should be developed in the early stages and with the participation of the provider(s), because their cooperation will be needed to execute such plans. Moreover, in the post September 11 world, contingency plans for disaster recovery should be fully developed and tested by the provider and/or buyer.
- *Developing a compelling stakeholder communication strategy.* Outsourcing should continue to be a sensitive political issue for the foreseeable future. Forrester Research estimates that over the next 12 years, 3.3 million US jobs, accounting for \$100 billion in wages, will move offshore. Such numbers will likely fuel the current backlash. To minimize reputational risks, companies should develop a well thought-out communication strategy with respect to their outsourcing initiatives. Besides communicating to external groups such as customers, unions, and governmental entities, internal communication is important given the potential for disgruntled employees to undermine outsourcing initiatives.

To develop, coordinate and implement project management controls and the above risk mitigation strategies, companies are putting in place centralized programme management offices (PMOs) as governance structures. These PMOs take a portfolio approach in allocating resources and monitoring vendor performance to ensure optimal performance. The PMOs also represent a centre of excellence for project management skills and resources, including sourcing and managing external resources such as consultants, tax experts and lawyers. Regardless of whether

a PMO is established, companies involved in, or planning to initiate, outsourcing arrangements should establish the appropriate operational risk controls discussed in the chapter. Otherwise, given the strategic importance of IT and outsourcing, the ‘next best thing’ might very well become the company’s worst nightmare.

References

Cumming, C M, and Hirtle, B J (2001) The challenges of risk management in diversified financial companies. *Federal Reserve Bank of New York Economic Policy Review*, March

Hernandez, J V, Sanchez, L M, and Ceske, R (2000) Quantifying event risk: the next convergence. *Journal of Risk Finance*, 1(3), pp. 9–23.

Lam, J (2003a) A unified management and capital framework for operational risk. *RMA Journal*, Feb., pp. 26–29.

Lam, J (2003b) *Enterprise Risk Management – from Incentives to Controls*. Hoboken, NJ: Wiley.

III.C.3 Operational Value-at-Risk

Carol Alexander¹

Many firms may wish to apply an ‘advanced measurement approach’ (AMA) to assess their capital to cover operational risk. Under the new Basel Accord that comes into force at the end of 2006, banks will at first apply a ‘top-down’ method (either the ‘basic indicator’ or ‘standardised’ approach) to assess their operational risk regulatory capital. However, by the end of 2007 they will be able to apply an AMA – and hopefully reduce their regulatory capital charge – provided they meet certain qualitative and quantitative criteria. Rating agencies are another driving force behind the implementation of AMA. Banks and corporates that aim for a high credit rating require an accurate assessment of their operational risks to convince rating agencies that capitalization is adequate. The ‘top-down’ methods provide only a crude estimate of operational risk, based on the unrealistic assumption that operational risks increase proportionally with gross income (see Section III.C.2.4.1).

Quantitative risk management requires an understanding of the ‘value-at-risk’ (VaR) models that are used to assess market, credit and operational risk capital. Operational VaR modelling is the subject of this chapter: Section III.C.3.1 outlines the ‘loss model’ approach to computing operational risk capital (ORC). Sections III.C.3.2 and III.C.3.3 examine how to apply some standard functional forms for the frequency distribution and severity distribution. Sections III.C.3.4 and III.C.3.5 describe how each component of ORC is estimated using (a) analytic and (b) simulation methods, and then Section III.C.3.6 explains how the component ORC estimates are aggregated over all business lines and event types to obtain the total ORC estimate for the firm. Section III.C.3.7 concludes

III.C.3.1 The ‘Loss Model’ Approach

The actuarial ‘loss model’ approach has recently become accepted by the industry as the generic AMA for the determination of operational risk regulatory capital for the new Basel 2 Accord (see Sections III.0.3 and III.C.1.7 for further details). Consequently, the loss model approach may also be favoured by rating agencies for firms of high credit quality. But even without this external pressure, many firms will want to adopt operational loss models as a key element of good risk management practice.

¹ Chair of Risk Management and Director of Research, ISMA Centre, Business School, University of Reading, UK.
Copyright © 2004 C. Alexander and The Professional Risk Managers’ International Association.

Prior to implementing an AMA the firm must identify events that are linked to operational risks, and map these events to an operational risk ‘matrix’ such as that based on the Basel 2 consultative documents (see Basel Committee on Banking Supervision, 2001) and which is shown in Table III.C.3.1. Each element in the matrix defines an operational risk ‘type’ by its business line and operational event category. In the AMA, operational risk capital is first assessed separately for each risk type for which the AMA is the designated approach.² Then the component estimates are aggregated to obtain the total AMA operational risk capital for the firm. The AMA could be chosen for only the most important risk types and this depends on the nature of the business: that is, the definition of the important event types and business lines will be specific to the firm’s operations. For example, operational losses for clearing and settlements firms may be concentrated in processing risks and systems risks.

Table III.C.3.1: The operational risk matrix

	Internal Fraud	External Fraud	Employment Practices & Workplace Safety	Clients, Products & Business Practices	Damage to Physical Assets	Business Disruption & System Failures	Execution, Delivery & Process Management
Corporate Finance							
Trading & Sales							
Retail Banking							
Commercial Banking							
Payment & Settlement							
Agency & Custody							
Asset Management							
Retail Brokerage							

The definition of business units and event types for the operational risk matrix can be specific to the firm. It will be natural to follow pre-existing internal definitions of business units and to define event types that capture the important operational risks. It should also take account of the granularity that is required for the AMA calculations. Increasing levels of granularity are necessary to include the impact of insurance cover, which may only be available for some event types in certain lines of business.

² Corporates are, of course, free to pick and choose which risk types they assess using AMA. Indeed, banks are also afforded some flexibility: under the new Basel Accord they can choose the AMA for some risk types and apply the standardized approach to others. However, once a risk type has been chosen for AMA modelling, the bank will not be allowed in future to apply a more basic risk capital assessment method.

It is desirable to isolate those elements of the matrix that are likely to be dominant in the final aggregation, as these risk types should be the main priority for risk control. Unfortunately, these can be precisely those risks for which the data are very subjective, consisting of expert opinions or risk self-assessments that have a large element of uncertainty. Somehow, this uncertainty in the data must be included in the risk model. Qualitative judgements must be translated into quantitative assessments of risk capital using appropriate statistical methodologies. Many firms now aim to do this through a ‘risk self-assessment’ process. A risk self-assessment gives a forward-looking, subjective estimate of the loss model parameters. Risk self-assessment programs can be facilitated in the same way as control self-assessments (see Section III.C.1.9).

Operational risks may be categorised in terms of:

- *frequency*, the number of loss events during a certain time period; and
- *severity*, the impact of the event in terms of financial loss.

Risks with very low frequency and high severity, such as a massive fraud or a terrorist attack, could jeopardise the whole future of the firm. These are the risks associated with losses that will lie in the very upper tail of the total loss distribution. Risk capital is not really designed to cover these risks. However, they might be insurable. Risks with high frequency and low severity, which include credit card fraud and processing risks, can have a high *expected loss* but will have relatively low *unexpected loss*. That is, the range of loss outcomes is relatively narrow. If expected losses for the high-frequency, low-severity risks are covered by the general provisions of the business, the implication is that ORC requirements for these risk types will be relatively low. If this is not the case, then expected losses should be included in the risk capital. Unless expected losses are very high, the risk capital will still be lower than that for medium-frequency, medium-severity risks. These latter risks are the legal risks, the minor frauds, the fines from improper practices, the large system failures and so forth. In general, these should be the main focus of the AMA.

An example of loss data is shown in Table III.C.3.2. For simplicity and only for the purposes of this illustration, the exact date of the loss is not actually recorded, only the quarter into which it falls. This allows one to classify data by quarterly frequency – or by semi-annual or annual frequency – but not by monthly or lower frequencies. Suppose we choose the quarterly period, so the frequency distribution will be of the number of loss events per quarter. First we ignore the severity data, and consider only the dates of the loss events. There were no quarters in which no loss events occurred, no quarters in which 1 loss event occurs, but there was one quarter in which 2 loss events occurred (2001:Q3) and two quarters in which 3 loss events occurred (2000:Q1 and 2000:Q3). Continuing counting in this way, we can draw an empirical frequency

density. Secondly, returning to the example loss data, we now ignore the date of loss, consider only the loss amounts and hence construct the empirical severity density.

Table III.C.3.2: Example of historical loss experience data

Date	Loss (£000)	Date	Loss (£000)	Date	Loss (£000)
2000:Q1	4.45	2001:Q1	7.51	2002:Q1	1.12
2000:Q1	13.08	2001:Q1	1.17	2002:Q1	4.06
2000:Q1	29.38	2001:Q1	1.35	2002:Q1	34.55
2000:Q2	25.92	2001:Q1	105.45	2002:Q1	10.24
2000:Q2	39.10	2001:Q1	37.24	2002:Q1	24.17
2000:Q2	12.92	2001:Q1	16.55	2002:Q1	11.01
2000:Q2	1.24	2001:Q2	7.34	2002:Q1	3.89
2000:Q3	8.01	2001:Q2	1.35	2002:Q1	187.50
2000:Q3	12.17	2001:Q2	1.50	2002:Q1	13.21
2000:Q3	13.88	2001:Q2	1.19	2002:Q1	4.49
2000:Q4	53.37	2001:Q2	2.80	2002:Q2	2.10
2000:Q4	5.89	2001:Q3	3.00	2002:Q2	2.20
2000:Q4	1.32	2001:Q3	6.82	2002:Q2	2.31
2000:Q4	7.11	2001:Q4	1.73	2002:Q2	25.00
		2001:Q4	231.65	2002:Q2	3.81
		2001:Q4	5.00	2002:Q3	1.48
		2001:Q4	3.10	2002:Q3	1.57
		2001:Q4	26.45	2002:Q3	20.33
		2001:Q4	12.62	2002:Q3	43.78
		2001:Q4	2.32	2002:Q3	3.62
		2001:Q4	71.12	2002:Q3	45.72
		2001:Q4	1.73	2002:Q4	142.59
				2002:Q4	20.73
				2002:Q4	31.96
				2002:Q4	55.60

For some real loss experience data based on a record going back over 24 months of operational loss events, Figure III.C.3.1 illustrates the empirical frequency density (the number of loss events per month). It shows that, out of the 24 months, there was only one month in which less than 10 loss events occurred; there were five months for which between 10 and 19 loss events occurred; three months for which between 20 and 29 loss events occurred, and so forth.

Figure III.C.3.1: Example of (monthly) frequency distribution

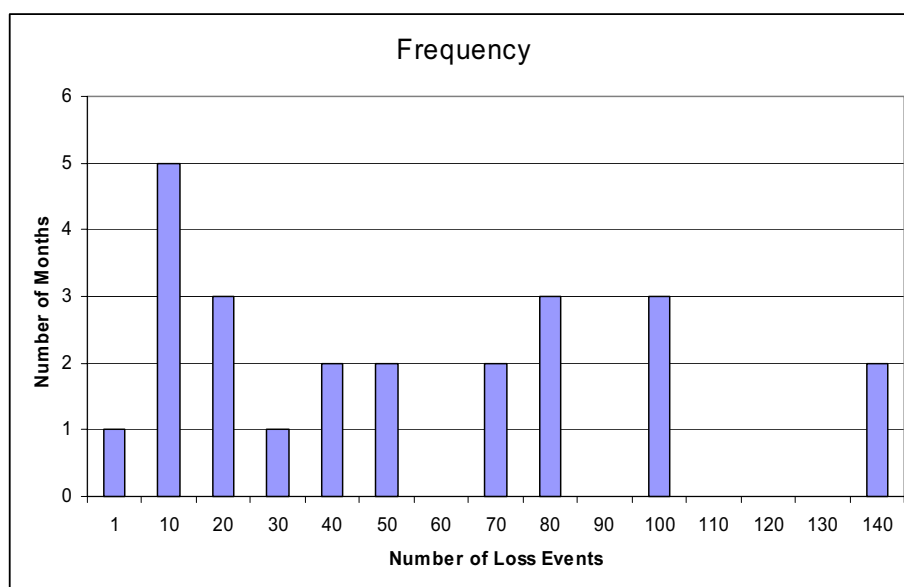
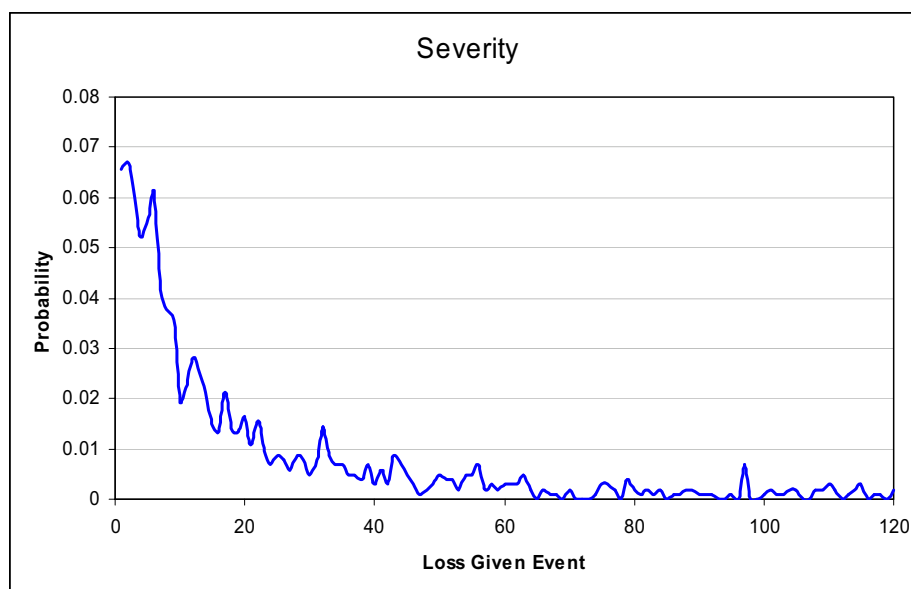


Figure III.C.3.2 illustrates the empirical severity density obtained from the same data. Only losses in excess of €1000 are recorded, so the severity data are truncated at the lower end. Special methods need to be applied to detruncate the severity data so that the full severity density can be recovered.

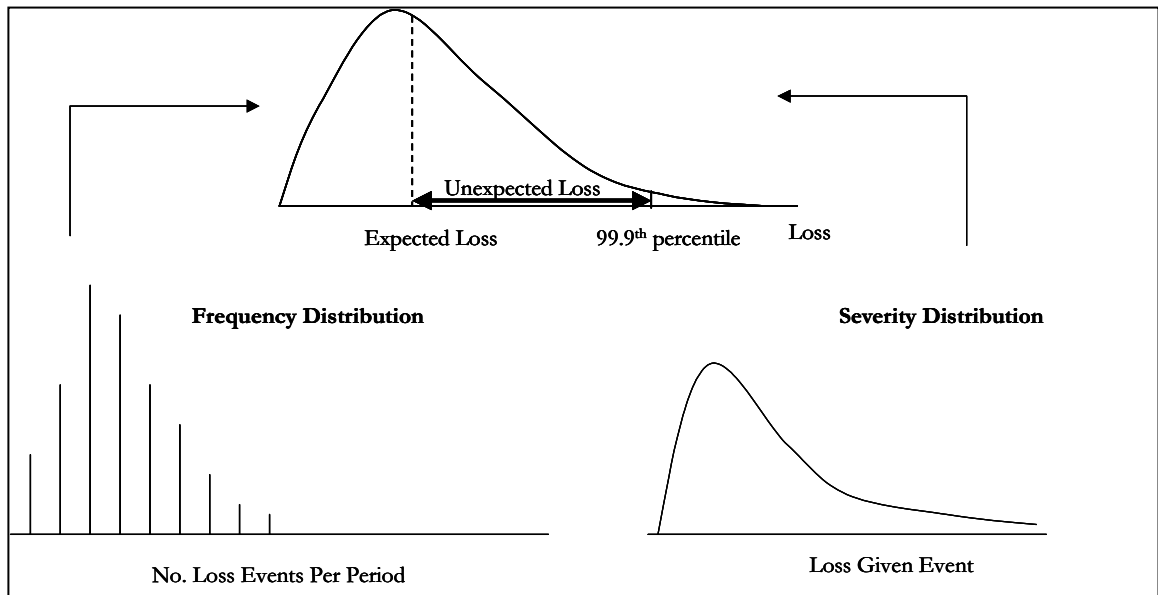
Figure III.C.3.2: Example of loss severity distribution



In summary, for a given operational risk type we construct a discrete frequency density $b(n)$ of the number of loss events n per period, and a continuous density $g(l)$ representing the loss severity, L . The density function for the *total loss* distribution $f(x)$ is then given by compounding these two densities, usually under the assumption that loss frequency and loss severity are independent. More details of the method for compounding these densities will be given in Section III.C.3.5 below.

Figure III.C.3.3 gives a diagrammatic representation of the relationship between the total loss distribution and its underlying frequency and severity distributions. The basic time period for the frequency density here is one year, and in this case the total loss distribution is also called the ‘annual’ loss distribution. This is a convenient terminology because later we need to aggregate several ‘total’ loss distributions over different risk types into a ‘total total’ loss distribution. However, by referring to each component as an ‘annual’ loss distribution (or a ‘quarterly’ loss distribution if the frequency is measured at quarterly intervals, or ‘monthly’ and so on) there is no confusion about what is meant by the ‘total’ loss distribution.

Figure III.C.3.3: Representation of the loss model



Marked on the density function for the total loss distribution are the

- *expected loss*, that is, the mean of this distribution; and the
- *unexpected loss at the 99.9th percentile*, that is, the unexpected loss at the α th percentile is the difference between the upper α th percentile and the mean of the annual loss distribution

ORC is held to cover all losses, other than highly exceptional losses, that are not already covered by the normal cost of the business. The definition of what one means by ‘highly exceptional’ translates into the definition of a percentile of the loss distribution, such that only losses exceeding this amount are ‘highly exceptional’. One should attempt to control these, using scenario analysis. As already mentioned, often expected losses are already included in the normal cost of business. For example, the expected loss from credit card fraud may be included in the balance sheet under ‘operating costs’. In that case ORC should cover only the unexpected loss at some predefined percentile of the loss distribution.

The ORC is thus defined using the VaR metric, for some risk horizon (defined below) and at some percentile. The Basel Committee recommend 99.9% and a one-year horizon for the calculation of ORC in banks; internally, for companies wishing to maintain a high credit rating, it is common to use an even higher percentile that is consistent with their desired credit rating. For instance, a firm that targets a AA rating will typically be measuring economic capital at the 99.97th percentile.

III.C.3.2 The Frequency Distribution

The total operational loss refers to a fixed time period over which these events are to be observed. This time period is usually called the *risk horizon* of the loss model to emphasise that it is a forward-looking time interval starting from today. For regulatory purposes, both operational and credit risk horizons are set at one year, but for internal purposes it is also common to use risk horizons of less than or more than one year. For ease of exposition we shall henceforth only refer to the one-year horizon – hence the AMA aims to model the *annual* loss distribution.

Having defined a risk horizon, the probability of a loss event, which has no time dimension, can be translated into the loss frequency, that is, the number of loss events occurring during the risk horizon. In particular, the expected loss frequency, denoted lambda (λ), is the product of the expected total number of events, N , during the risk horizon, including events for which no operational loss was made, and the expected loss probability, p :

$$\lambda = Np. \quad (\text{III.C.3.1})$$

Sometimes it is convenient to forecast λ directly – this is the case when we cannot quantify the total number of events N and we only observe loss events – and in other cases it is best to forecast N and p separately – for example, N could be the target number of transactions over the next year, and in that case it is p , not λ , that we should attempt to forecast using loss experience and/or risk self-assessment data.

Loss frequency (the number of loss events occurring during the risk horizon) is a discrete random variable: it can only take the values 0, 1, 2, ... , N . A fundamental density for such a discrete random variable (which nevertheless is only appropriate under certain assumptions)³ is the well-known *binomial* density (see Section II.E.4.1),

$$b(n) = \binom{N}{n} p^n (1-p)^{N-n} \quad n = 0, 1, \dots, N. \quad (\text{III.C.3.2})$$

The binomial frequency is only used when a value for N can be specified. For example,

- for Trading & Sales/Client, Products and Business Practice Risk, N could be the target number of deals over the next year;
- for Retail Banking/External Fraud, N could be the total number of credit cards in issuance during the risk horizon.⁴

³ We must assume that the probability of a loss event is the same for all the events in this risk type, and therefore equal to p ; and that operational events are independent of each other.

⁴ Note that here N would be so large that the Poisson distribution can be used instead of the binomial distribution. However, firms may still wish to employ the binomial distribution when, for instance, their target for N over the next year is quite different from the historical value of N .

It is not always possible to specify N . However, since p is normally small the binomial distribution can often be well approximated by the *Poisson* distribution, which has the single parameter λ , the expected frequency, as in (III.C.3.1) above. Incidentally, λ is also equal to the variance of the Poisson distribution. The Poisson distribution has the density function (see Section II.E.4.2)

$$b(n) = \frac{\lambda^n \exp(-\lambda)}{n!} \quad n = 0, 1, 2, \dots \quad (\text{III.C.3.3})$$

If the empirical frequency density is not well modelled by a Poisson distribution – for example, one could equate the mean frequency observed empirically with λ , but then find that the sample variance is significantly different from λ – an alternative, more flexible functional form is the *negative binomial* distribution, with density function

$$b(n) = \binom{\alpha + n - 1}{n} \left(\frac{1}{1 + \beta} \right)^\alpha \left(\frac{\beta}{1 + \beta} \right)^n \quad n = 0, 1, 2, \dots, \quad (\text{III.C.3.4})$$

which has mean $\alpha\beta$ and variance $\alpha\beta^2$.

There is absolutely no point in applying a statistical test to decide which of the frequency distributions provides the closest fit to loss data: the binomial is only applicable when a ‘number of events’ can be quantified (and is small) and the negative binomial has two parameters so it will always fit better than the Poisson. However, this does not imply that one should always choose the negative binomial as the frequency functional form. In contrast to market and credit risk, for operational risk the precise fitting of data by choosing the best functional form is *not* a main source of model risk. In fact the model risk arising from inappropriate and/or *ad hoc* methods when handling the data is a much more important source of operational VaR model risk.⁵

The choice of functional form for the frequency distribution should depend on both the type of data and the source(s) of the data – internal and/or external loss data and/or risk self-assessments. It is very difficult to design psychologically meaningful questions in a risk self-assessment that are compatible with a negative binomial frequency. On the other hand, the question ‘what is the expected number of loss events next year?’ will directly invoke the parameter λ for the Poisson distribution. Also, it is difficult to apply the binomial distribution to external consortium data because the consortium will not normally be recording a value of N for each bank. These restrictions imply that it is often the Poisson distribution that is used in practice.

⁵ But of course, in common with market and credit risk, the main model risk is ‘aggregation risk’. This stems from making inappropriate assumptions regarding dependencies when aggregating risks. This has by far the most influence on the total VaR estimate. See, for instance, Chapter III.A.3.

Example III.C.3.1: Estimating a Poisson frequency from historical data

(i) *High-frequency risk type.* Suppose historical loss events give the following data on just the number of loss events recorded each month, over the last 2 years:

Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Number of Loss Events	20	13	24	26	25	21	17	13	21	30	16	24	31	20	19	21	14	14	15	18	16	21	22	19

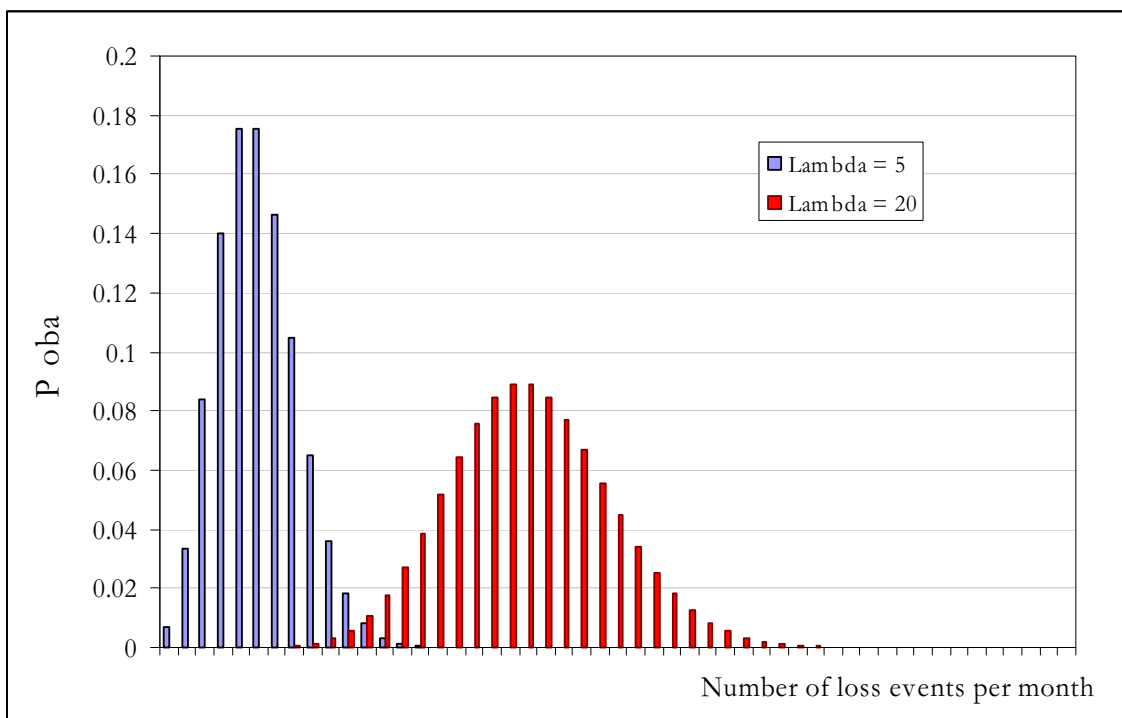
The total number of loss events recorded was 480, so the average number of loss events per month is 20. The monthly frequency distribution is therefore estimated as a Poisson distribution with $\lambda = 20$. The density function is shown in red in Figure III.C.3.4.

(ii) *Low(er)-frequency risk type.* Suppose historical loss events give the following data on just the number of loss events recorded each month, over the last 2 years:

Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Number of Loss Events	0	12	3	0	8	4	10	1	0	9	2	10	3	5	3	5	1	7	4	7	10	5	7	4

The total number of loss events recorded is now only 120, so the average number of loss events per month is 5. The monthly frequency distribution is therefore estimated as a Poisson distribution with $\lambda = 5$. The density function is shown in blue in Figure III.C.3.4.

Figure III.C.3.4: Poisson frequency densities for high-frequency and low-frequency risks



Notes:

1. The Poisson annual frequency density is obtained by multiplying the estimated lambda from monthly data (20 and 5 in the above example) by 12.
2. Lower-frequency risks have more skewed and leptokurtic frequency densities than high-frequency risks. This property influences the compound distribution, so that the annual loss distribution will also be highly skewed and leptokurtic for the low-frequency, high-severity risks.

III.C.3.3 The Severity Distribution

Now consider how to ‘fit’ a severity distribution, such as that shown in Figure III.C.3.2. Various functional forms are available for continuous random variables like severity, as described in Chapter II.E. The *lognormal* distribution for loss severity, L , has the density function:

$$g(l) = \frac{1}{\sqrt{2\pi}\sigma l} \exp\left(-\frac{1}{2}\left(\frac{\ln l - \mu}{\sigma}\right)^2\right) \quad (l > 0). \quad \text{(III.C.3.5)}$$

Here the log of the loss severity (or *log severity* for short, denoted $\ln L$) is normally distributed with mean μ and variance σ^2 . This is a very common distribution in financial mathematics, and more details are given in Section II.E.4.5.

High-frequency risks can have severity distributions that are relatively lognormal, but low-frequency risks can have severity distributions that are too skewed and leptokurtic to be well captured by the lognormal density function. Common choices, therefore, are the gamma density,

$$g(l) = \frac{l^{\alpha-1} \exp(-l/\beta)}{\beta^\alpha \Gamma(\alpha)} \quad (l > 0), \quad \text{(III.C.3.6)}$$

where $\Gamma(\cdot)$ denotes the gamma function, and the two-parameter hyperbolic density,

$$g(l) = \frac{\exp(-\alpha\sqrt{\beta^2 + l^2})}{2\beta B(\alpha\beta)} \quad (l > 0), \quad \text{(III.C.3.7)}$$

where $B(\cdot)$ denotes the Bessel function of the first kind. On first sight these may look daunting, but Excel does have in-built functions for these common distributions (with the obvious names). Other functional forms for the severity that have been considered by some banks include the generalised hyperbolic, lognormal mixtures, and general mixture distributions.

Again, there is absolutely no point in applying a statistical test to decide which of these frequency distributions provides the closest fit to loss data: the generalised four-parameter hyperbolic distribution will always fit best. However, this does not imply that one should always choose the generalised hyperbolic distribution as the severity functional form. Again, this choice again depends on both the type of data and the source(s) of the data – internal and/or external loss data and/or risk self-assessments. For example, it is very difficult to design a risk self-assessment that is compatible with any severity density having more than two parameters.

In some databases, for instance those constructed from public, ‘newsworthy’ events, only very extreme losses are recorded. There is an implicit high threshold for the losses included in the database and thus some analysts have attempted to model the severity of these losses using the generalised Pareto and other distributions from the class of distributions in ‘extreme-value theory’ (EVT). Whilst EVT may have found useful applications to high-frequency, tic-by-tic financial market data, one should not forget that it was introduced (almost 50 years ago) to model the distributions of extreme values in repetitive, independent and identically distributed processes, such as those observed in the physical sciences; see, for instance, Gumbel (1958) and Embrechts *et al.* (1991). To attempt to fit a generalised Pareto distribution, or any other extreme-value distribution, to the sparse and fragmented data that are available for very large operational losses is, in my view, a triumph of hope over reason. Nevertheless, some AMAs are currently attempting to incorporate extreme-value densities, and readers should therefore have some understanding of them.

The ‘peaks-over-threshold’ (POT) model applies when losses over a high and predefined threshold u are recorded. The distribution function G_u of the excess losses, $X - u$, has a simple relation to the distribution $F(x)$ of the loss severity X . In fact

$$G_u(y) = \text{prob}(X - u < y \mid X > u) = [F(y + u) - F(u)] / [1 - F(u)]. \quad (\text{III.C.3.8})$$

For many choices of underlying distribution $F(x)$ the distribution $G_u(y)$ will belong to the class of *generalised Pareto distributions* (GPDs) given by:

$$G_u(y) = \begin{cases} 1 - \exp(-y/\beta) & \text{if } \xi = 0, \\ 1 - (1 + \xi y/\beta)^{-1/\xi} & \text{if } \xi \neq 0. \end{cases} \quad (\text{III.C.3.9})$$

The parameters β and ξ will depend on the type of underlying loss distribution $F(x)$ and on the choice of threshold u . Some generalised Pareto densities for different values of β and ξ are shown in Figures III.C.3.5 and III.C.3.6. Note that the primary effect of increasing β is to increase the

range of the density (Figure III.C.3.5), and that ξ is called the ‘tail index’ precisely because as ξ increases so does the weight in the tails of the GPD (see Figure III.C.3.6).

Figure III.C.3.5: Generalised Pareto densities ($\xi = 1$)

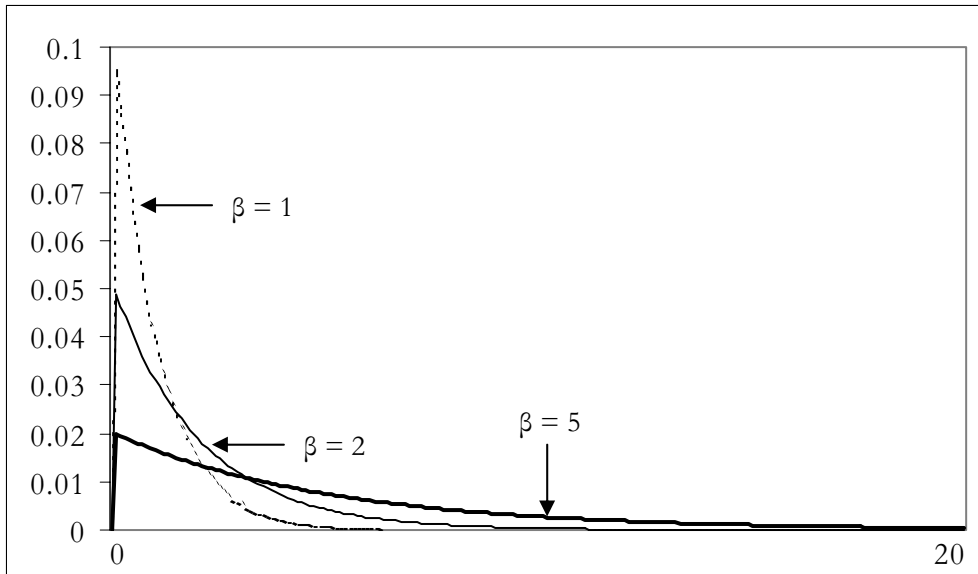
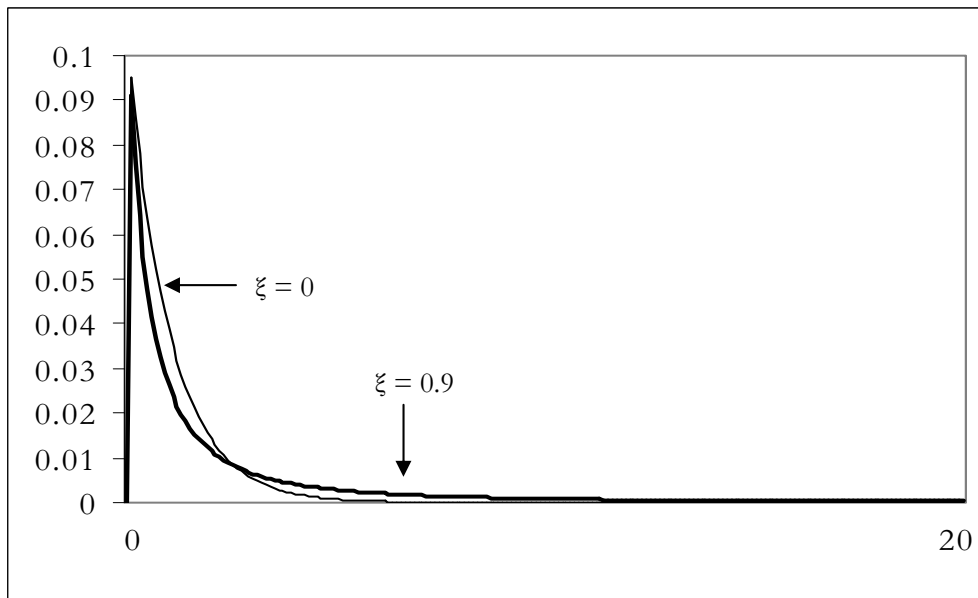


Figure III.C.3.6: Generalised Pareto densities ($\beta = 1$)



III.C.3.4 The Internal Measurement Approach

Under certain assumptions (which are rather strong, but nevertheless appear to be admissible by regulators), there are simple analytic formulae for the expected loss and the unexpected loss in the annual loss distribution. These formulae are based on what the Basel Committee has called the ‘internal measurement approach’ (IMA). The basic formula for the IMA risk capital calculation given in the proposed Basel 2 Accord is:

$$ORC = \textit{gamma} \times \textit{expected annual loss} = \gamma \times NpL, \quad (\text{III.C.3.10})$$

where N is a volume indicator (a proxy for the number of operational events), p is the expected probability of a loss event, L is the loss given event, and γ , ‘gamma’, is a multiplier that depends on the operational risk type.

Note that NpL only corresponds to the expected annual loss when the loss frequency is binomially distributed and the loss severity is not regarded as a random variable. More generally, with a Poisson frequency distribution having expected frequency $\lambda (= Np)$:

$$ORC = \gamma \times \lambda \times L. \quad (\text{III.C.3.11})$$

Note that a very strong assumption of the IMA is that *each time a loss is incurred, exactly the same amount is lost* (within a given risk type). Introduction of severity uncertainty, as in the ‘loss distribution approach’ (LDA) described below, will always increase the ORC, often by a factor of 5 or more. Thus the IMA provides only a useful benchmark, a lower bound for the operational risk capital calculated using the full simulation method that we shall describe presently.

I have shown elsewhere (Alexander, 2003, p. 151) how the value of γ in (III.C.3.11) can be calculated and provide statistical tables for γ : it only depends on the percentile and the expected loss frequency. Instead of following Consultative Paper 2.5 (Basel Committee, 2001) and writing unexpected loss as a multiple (γ) of expected loss, I write unexpected loss as a multiple phi (φ) of the loss standard deviation. That is,

$$ORC = \varphi \times \textit{standard deviation of annual loss}. \quad (\text{III.C.3.12})$$

Recall that ORC is measured by a VaR metric, and is *either* unexpected loss at the 99.9th percentile *or* the 99.9th percentile itself, the latter being the case when expected losses are not already provisioned. Thus (III.C.3.12) can be rewritten as:

$$\left. \begin{aligned} \varphi &= [(99.9\text{-ile} - \text{mean}) / \text{standard deviation}] && \text{if expected losses are provisioned,} \\ \varphi &= [(99.9\text{-ile} / \text{standard deviation})] && \text{otherwise.} \end{aligned} \right\} \text{(III.C.3.13)}$$

This formulation shows that, once we have estimated λ (either from a risk self-assessment or from loss data, as explained in Section III.C.3.2) it is easy to calibrate φ . Because loss severity is assumed to be the *same* amount, L , every time a loss is made, all the randomness in the annual loss distribution comes only from the frequency distribution. Hence all quantities on the right-hand side of (III.C.3.13) – the 99.9th percentile, the mean and the standard deviation – are just L times the respective quantities of the frequency density. Thus L cancels, and the right-hand side of (III.C.3.13) really just refers to the 99.9th percentile, the mean and the standard deviation of the frequency density.

Having obtained φ we can obtain γ as follows: compare (III.C.3.10) and (III.C.3.12). Since ORC is the same in both, equating them gives:

$$\gamma \times \text{expected annual loss} = \varphi \times \text{standard deviation of annual loss}$$

so γ is a multiple of φ : in fact

$$\gamma = \varphi \times (\text{standard deviation} / \text{mean}).$$

The same argument as above (i.e. that L cancels) implies that γ and φ are related through a factor (*standard deviation/mean*) of the frequency density. For instance, when the frequency is Poisson, which has mean and variance λ ,

$$\gamma = \varphi / \sqrt{\lambda}. \tag{III.C.3.14}$$

In this way, we construct statistical tables for the *gamma* factor in the IMA, a different table for each choice of frequency distribution. Table III.C.3.3 gives the Poisson gamma tables (φ is also shown).

Table III.C.3.3: The ‘gamma’ in the IMA (Poisson frequency)

$\lambda \rightarrow$	100	50	40	30	20	10
99.9%-ile	131.805	72.751	60.452	47.812	34.714	20.662
φ	3.180	3.218	3.234	3.252	3.290	3.372
γ	0.318	0.455	0.511	0.594	0.736	1.066
$\lambda \rightarrow$	8	6	5	4	3	2
99.9%-ile	17.630	14.449	12.771	10.956	9.127	7.113
φ	3.405	3.449	3.475	3.478	3.537	3.615
γ	1.204	1.408	1.554	1.739	2.042	2.556
$\lambda \rightarrow$	1	0.9	0.8	0.7	0.6	0.5
99.9%-ile	4.868	4.551	4.234	3.914	3.584	3.255
φ	3.868	3.848	3.839	3.841	3.853	3.896
γ	3.868	4.056	4.292	4.591	4.974	5.510
$\lambda \rightarrow$	0.4	0.3	0.2	0.1	0.05	0.01
99.9%-ile	2.908	2.490	2.072	1.421	1.065	0.904
φ	3.965	3.998	4.187	4.176	4.541	8.940
γ	6.269	7.300	9.362	13.205	20.306	89.401

Source: Alexander (2003). Table reproduced by kind permission of Pearson Education (Financial Times-Prentice Hall).

For instance, when $\lambda = 100$ (i.e. a very high-frequency risk) the 99.9th percentile of the Poisson distribution is 131.805. Thus, for a capital charge based only on unexpected loss, not on expected loss, we have

$$\varphi = (131.805 - 100) / \sqrt{100} = 31.805/10 = 3.1805,$$

$$\gamma = 3.1805 / \sqrt{100} = 0.31805.$$

Note that φ does not change much with λ , but that γ does. For low-frequency risks γ is very large indeed, but for high-frequency risks it is very low. As λ increases above 100, for very high-frequency risks, the frequency distribution approaches the normal distribution so φ tends to a lower limit of 3.09 (this is the 99.9th percentile in the standard normal distribution); however, γ tends to a lower limit of zero!

Also note that the ORC should increase as the *square root* of the expected frequency. For example, in the Poisson frequency, combining (III.C.3.11) with (III.C.3.14) gives $ORC = \varphi \times \sqrt{\lambda} \times L$, and we know that φ is around 3 or 4, unless the expected number of loss events is less than 1 every 50 years. So in the AMA the ORC will *not* be linearly related to the size of the bank’s operations,

as it is under the basic indicator or standardised approach. Hence doubling the size of one's operations should only lead to an increase of $\sqrt{2}$ in the capital charge; this may be an incentive for banks to use the AMA when they are permitted to do so at the end of 2007. The ORC also is linearly related to loss severity – yet another reason why high-severity risks (which are by definition also low-frequency) will attract higher capital charges than low-severity risks; the following two examples illustrate this point.

Example III.C.3.2: Credit card operational risk

Question:

50,000 credit cards will be in issuance during the forthcoming year and the expected probability of a credit card fraud on any card is 0.05. If every fraud gives rise to an operational loss of €1000, calculate (a) the expected loss, (b) the ORC at the 99.9th percentile, with and without the assumption that the expected loss is already covered by the normal cost of the business.

Answer:

(a) $\lambda = 50,000 \times 0.05 = 2500$ so expected loss = $2500 \times 1000 = \text{€}2.5$ million.

(b) For such a high λ we have $\varphi = 3.1$ and ORC is estimated using (III.C.2.10).

(c) Standard deviation of loss is $\sqrt{\lambda L} = \sqrt{2500 \times 1000} = \text{€}50,000$ and so

$$\text{ORC} = 3.1 \times 50,000 = \text{€}155,000 \text{ (assuming expected loss is provisioned for),}$$

$$\text{ORC} = 155,000 + 2.5 \text{ million} = \text{€}2.655\text{m (if expected loss must be included).}$$

This example shows that for high-frequency, low-severity operational risks the expected loss is much greater than the unexpected loss. The expected losses are normally already provisioned in the normal cost of being in the credit card business. In that case the ORC is there to cover the *unexpected* losses. The ORC for this type of risk will therefore be very small, and this risk type will have little impact on the total ORC for a retail bank. Indeed, the next example shows that the total ORC will be dominated by the low-frequency, high-severity operational risks:

Example III.C.3.3: Transactions errors versus internal fraud

Show that the expected loss is the same in the following two cases, but that the ORC is far greater for the internal fraud:

Case A: 30,000 transactions are processed in the back office and the expected probability of a human error giving rise to an operational loss is 0.01. Each time a loss occurs the operational loss is €1000.

Case B: 60 deals are made in Corporate Finance and the probability of internal fraud resulting in any one of these making an operational loss is 0.0005. However, if such a loss is made, it will amount to €10m.

The ORC calculations based on (III.C.2.9) in Table III.C.3.4 below assume that expected loss is provisioned for; clearly, the same basic observation holds whether or not the additional sum of

€300,000 is added to the ORC figures. That is, even if the ORC estimates were €0.35369m and €7.22474m respectively, the low-frequency, high-severity risk type is the one that totally dominates the total ORC.

Table III.C.3.4: Results for Example III.C.3.3

	Case A	Case B
N	30,000	60
p	0.01	0.0005
L	1,000	10,000,000
Expected Loss (NpL)	300,000	300,000
Standard Deviation ($= \sqrt{(Np)L}$)	17,320.51	1,732,051
$\lambda (= Np)$	300	0.03
φ	3.1	3.998
$\gamma (= \varphi/\sqrt{\lambda})$	0.178979	23.08246
ORC (€m)	0.05369	6.92474

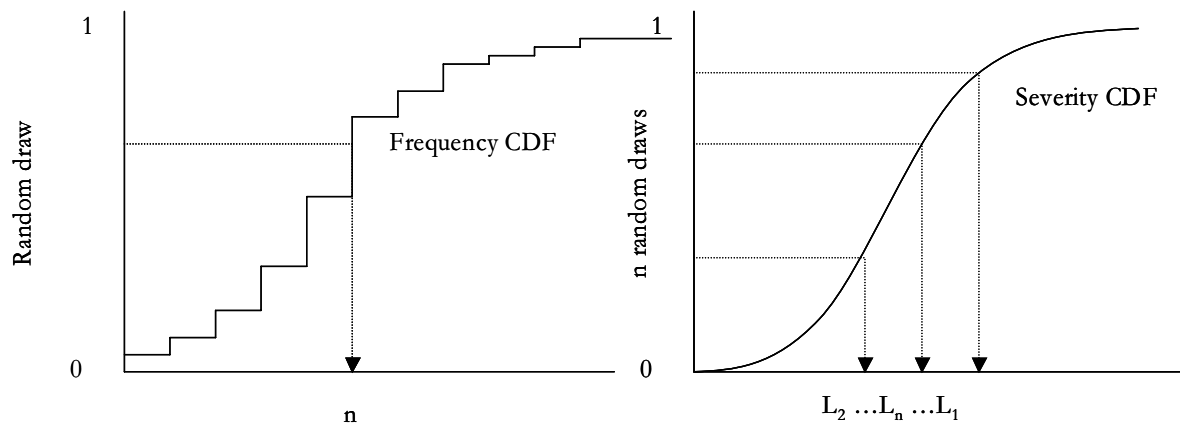
III.C.3.5 The Loss Distribution Approach

The standard assumption that frequency and severity are *independent* may not be very realistic, but it makes the construction of the compound distribution very simple. Monte Carlo simulation (see Section II.E.4.3) is used to generate the total loss distribution as the compound of the frequency and severity distribution. The simulation algorithm is as follows:

1. Take a random draw from the frequency distribution: suppose this simulates n loss events per period.
2. Take n random draws from the severity distribution: denote these simulated losses by L_1, L_2, \dots, L_n .
3. Sum the n simulated losses to obtain a total loss $X = L_1 + L_2 + \dots + L_n$.
4. Return to step 1, and repeat several thousand times: thus obtain X_1, \dots, X_M where the number of simulations M is very large.
5. Form the histogram of X_1, \dots, X_M : this represents the simulated total loss distribution.
6. The ORC for this risk type is then the difference between the 99.9th percentile and the mean of the simulated annual loss distribution, or just the 99.9th percentile, if expected losses are not provisioned for elsewhere.

Figure III.C.3.7 illustrates the first two steps in the simulation algorithm.⁶

Figure III.C.3.7: Simulating the annual loss distribution



$$\text{Total annual loss from one simulation} = \sum L_i = \text{TL}$$

Repeat to obtain $\text{TL}_1, \dots, \text{TL}_{10000}$

$$\text{ORC Estimate} = \text{Percentile}(\text{TL}_1, \dots, \text{TL}_{10000}, 0.999) - \text{Average}(\text{TL}_1, \dots, \text{TL}_{10000})$$

Elsewhere (Alexander, 2003, Ch. 7) I have compared the LDA estimate of ORC with the result of applying the IMA formula of Section III.C.3.4. I considered two cases:

- Without an assumption of loss severity uncertainty in the LDA, the result will be identical to the IMA estimate, provided enough simulations are used in the LDA calculation.
- The IMA estimate can be modified to include the assumption of loss severity uncertainty, in which case it is multiplied by the factor $\sqrt{1 + r^2}$, where

$$r = (\text{severity standard deviation} / \text{severity mean}). \quad (\text{III.C.3.15})$$

The resulting ‘modified’ IMA estimate is similar to the LDA estimate.

This shows that *operational risk capital will increase more or less in proportion with the severity standard deviation*. So when loss severity becomes more uncertain, this has a direct, almost linear impact on the capital charge. A very strong conclusion can be drawn. That is, for any given risk type: the LDA ORC is always *far* greater than the ORC calculated from an analytic formula based on the assumption that severity is non-random (and this includes the IMA formula). This difference will be most pronounced for operational risks where the loss severity is highly uncertain (so that r in

⁶ The use of empirical frequency and severity distributions is not advised, even if sufficient data are available to generate these distributions. There are two reasons for this. Firstly, the simulated annual loss distribution will not be an accurate representation if the same frequencies and severities are repeatedly sampled. Secondly, there will be no ability to carry out scenario analysis in the model unless one specifies and fits the parameters of a functional form for the severity and frequency distributions.

(III.C.3.15) is large): for these risk types, the risk capital calculation based on the LDA can easily give an ORC estimate that is 10 or 20 times larger than the IMA estimate.

III.C.3.6 Aggregating ORC

Having estimated the ORC for each ‘risk type’ for which the AMA is the designated approach, we must now ‘add up’ these ORC estimates to obtain the *total* ORC for the firm. The aggregation of risks – not just operational risks – is currently a hot topic of research. It is very difficult to aggregate risks (a) when they are assessed using a VaR metric at an early stage;⁷ (b) because the ‘total’ we get for the risks is very much influenced by the assumptions we make about the dependencies between the risks; and (c) because dependencies between risks are very difficult to assess. Thus risk aggregation is very difficult even within market and credit risks – it is a *very* thorny issue in operational risk!

Regarding dependency assumptions, the Basel Committee (2001) has stated: ‘The bank will be permitted to recognize empirical correlations in operational risk losses across business lines and event types, provided that it can demonstrate that its systems for measuring correlations are sound and implemented with integrity’.

Dependencies between operational risks are common.⁸ Indeed, they occur whenever two operational risk types share a common key risk driver (see Section III.C.2.3). For instance, risk drivers associated with ‘human’ risks – such as pay, training, management, workload – affect many types of operational risks, including employment practices, transactions processing, legal risks, and fraud.

Often, when aggregating risks, particularly when the VaR metric is applied, banks will make just two simple assumptions:

- *Full dependency*: This implies risks should simply be added to obtain the total risk; this is an approximate upper bound for the total risk and is often used for regulatory risk capital calculations (to err on the conservative side).
- *No dependency*: This is the assumption of ‘independence’. It implies that the total risk is the square root of the sum of the squares of the component risks. For aggregating different types of operational risks, this will give an approximate lower bound for the

⁷ Unlike the ‘variance’ risk metric, percentiles obey no simple standard rules. It is easy to aggregate variances (and to allow for correlation in the process) but it is not so simple to aggregate VaR.

⁸ I prefer not to use the term ‘correlation’ – instead I use the term ‘dependency’. Operational risks are likely to be dependent, in the sense that if one changes then so will the other, but they are not ‘correlated’. Correlation is a metric that refers to ‘jointly stationary’ random variables with elliptical distributions like the multivariate normal (see Section II.E.4.7) but there is no evidence that operational losses behave in this way.

total risk capital since it is unlikely that there will be many large *negative* dependencies between different operational risks.

The next example shows that small changes in ‘correlation’ between operational risk types will have a *huge* effect on the total risk estimate: for example, the total risk capital based on the aggregation of only two operational risk types can easily be doubled – or halved – depending on the assumption made about their dependency.

Example III.C.3.4: Effect of ‘correlation’ on risk aggregation

Consider the two annual loss distributions with density functions shown in Figure III.C.3.8. Figure III.C.3.9 shows the total loss under correlations of $\rho = 0.5, 0, -0.5$, respectively. Then Table III.C.3.5 shows that the expected loss is hardly affected by the assumptions made about co-dependencies of these two risks: it is approximately 22.4 in each case. However the unexpected loss at the 99.9th percentile (and at the 99th percentile) is very much affected by the assumption one makes about dependency.

Table III.C.3.5: Risk capital estimates under different correlation assumptions

	$\rho = -0.5$	$\rho = 0$	$\rho = 0.5$
Expected Loss	22.3909	22.3951	22.3977
99.9th Percentile	41.7658	48.7665	54.1660
Unexpected Loss	19.3749	26.3714	31.7683

Of course, the values of the correlation parameter were chosen arbitrarily in Example III.C.3.4. But it has shown that small changes in correlation can produce estimates of total operational risk capital that is doubled – or halved – even when aggregating only two annual loss distributions. Obviously the effect of correlation assumptions on the aggregation of many annual loss distributions to the total annual loss for the firm will be quite enormous.

Figure III.C.3.8: Two annual loss densities⁹

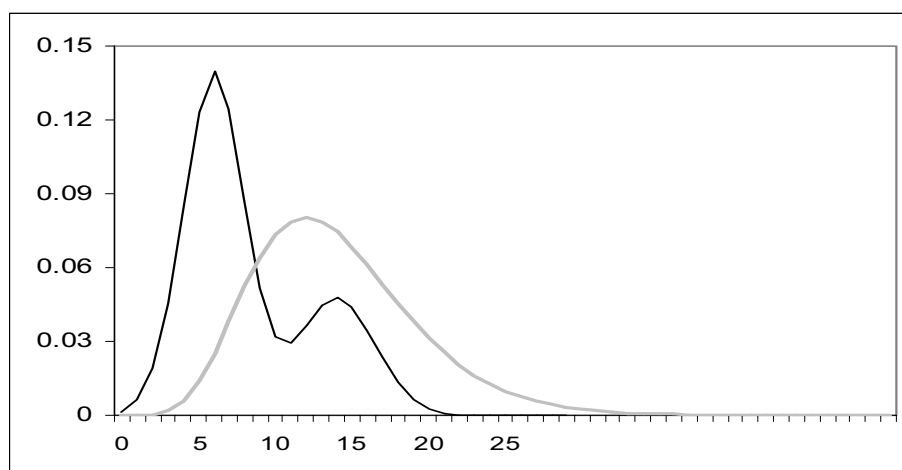
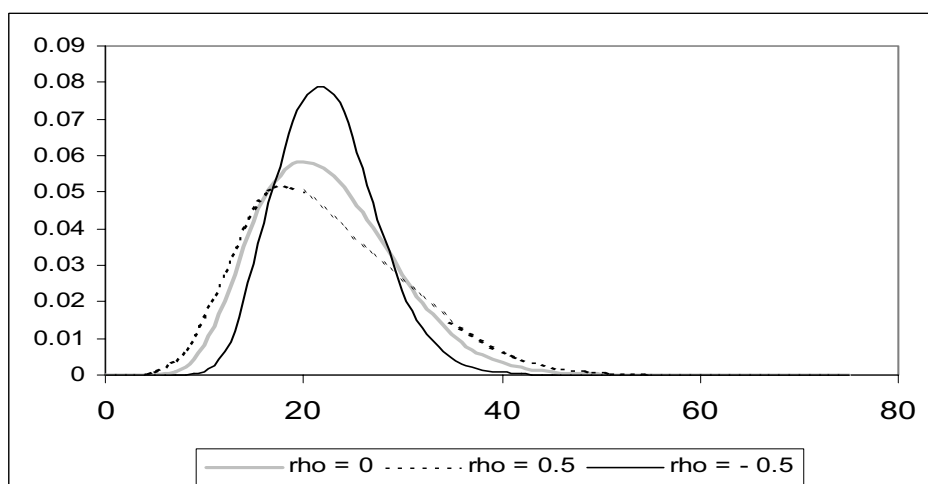


Figure III.C.3.9: The total loss distribution under different assumptions for correlation



III.C.3.7 Concluding Remarks

Operational risks often show a positive dependency because they experience the same directional influence from *common* key risk drivers. These drivers can be linked to ‘human’ risks, ‘systems’ risks and so forth. Thus the total risk estimate will be somewhere between a lower bound given by the square root of the sum of the squared risk estimates and an upper bound given by the sum of the risk estimates. These two bounds are usually far apart and it is very difficult to say exactly where, within these bounds, the total risk will be. Hence ‘aggregation risk’ is by far the most important source of model risk in an operational VaR model.

⁹ The bimodal density has been fitted by a mixture of two normal densities: with probability 0.3 the normal has mean 14 and standard deviation 2.5 and with probability 0.7 the normal has mean 6 and standard deviation 2. The other annual loss is gamma distributed with $\alpha = 7$ and $\beta = 2$.

The next most important source of model risk is the way that data are obtained and, subsequently, how these data are handled. The most important risk types for a *risk* estimate (as opposed to an expected loss estimate) are the low-frequency, high-severity risk types, such as internal fraud, acts of God (such as earthquakes), massive legal suits and so forth. Internal historical loss data on these risk types are simply not available as the frequency is just too low. The only way to include these risk types in the risk estimate is to use subjective data: such as a risk self-assessment, or external data from public sources, or from a data consortium. But risk self-assessment questionnaires require careful design and even more careful control of the responses, and external data need processing using *proper* statistical methods.

When data are subjective the proper statistical methods to use are Bayesian methods (see Chapter II.E). However, at the moment we are witnessing attempts to apply some of the traditional ‘classical’ methods that have been developed in market risk to operational risk analysis, as if the sparse and unreliable data on operational losses should be treated like the plentiful and reliable data on market prices. Operational risks are not at all like market or credit risks – operational risk analysis is about human behaviour, not market or firm behaviour! Thus we really should *not* be talking about ‘correlations’, or ‘tail distributions’ with data such as these. And software consultants who focus on fitting the ‘best’ functional form to operational risk data are missing the big picture (and making a fast buck in the process!). Getting the right assumptions about functional forms may be an important source of model risk in market and credit risk analysis but this is really not an important issue in operational risk assessment. The important issues are how to model dependencies and how to obtain reliable data. We may indeed apply a VaR metric to an operational loss distribution, but that does not mean that operational risk analysis is a statistical science with proper economic foundations, like market and credit risk. It is not. Operational risk is a behavioural science.

References

- Alexander, C (ed.) (2003) *Operational Risk: Regulation, Analysis and Management*. London: FT-Prentice Hall (Pearson Education).
- Basle Committee on Banking Supervision (2001) *Working Paper on the Regulatory Treatment of Operational Risk*, Consultative Paper 2.5, September., available from www.bis.org
- Gumbel, E J (1958) *Statistics of Extremes*. New York: Columbia University Press.
- Embrechts, P, Klüppelberg, C, and Mikosch, T (1991) *Modeling Extremal Events*. Berlin: Springer-Verlag.

III.0 Capital Allocation and RAPM

Andrew Aziz and Dan Rosen¹

III.0.1 Introduction

We introduce in this chapter the definitions and key concepts regarding capital, focusing on the important role that capital plays in financial institutions. Readers should already be familiar with Chapter I.A.5, which has presented the basic principles behind the capital structure of the firm. There it was argued that the *actual* capital – the physical capital that a firm holds – should be distinguished from the *optimal* level of capital. The optimal capital depends on many things, including capital *targets* that are associated with a desired level of ‘capital adequacy’ to cover the potential for losses made by the firm. Capital adequacy is assessed using internal models (for *economic* capital) and for banks a certain level of capital is imposed by external standards (*regulatory* capital). Capital adequacy is a measure of a firm’s ability to remain a going concern. Thus, targeted levels of capital are direct functions of the riskiness of the business activities or, from a balance sheet perspective, the riskiness of the assets.

In this chapter you will learn:

- the role of capital in financial institutions and the different types of capital;
- the key concepts and objectives behind regulatory capital, as well as the main calculations principles in the Basel I Accord and the current Basel II Accord;
- the definition and mechanics of economic capital as well as the methods to calculate it;
- the use of economic capital as a management tool for risk aggregation, risk-adjusted performance measurement and optimal decision making through capital allocation.

This introductory section presents the definition of capital and its role in financial institutions. We make the distinction between the various types of capital: book capital, economic capital and regulatory capital. We discuss briefly the use of economic capital as a management tool for risk aggregation, decision making and performance measurement.

III.0.1.1 Role of Capital in Financial Institution

Banks generate revenue by taking on exposure to their customers and by earning appropriate returns to compensate for the risk of this exposure. In general, if a bank takes on more risk, it can expect to earn a greater return. The trade-off, however, is that the same bank will, in general, increase the possibility of facing losses to the extent that it defaults on its debt obligations and is

¹ Algorithmics Inc.

forced out of business. Banks that are managed well will attempt to maximise their returns only through risk taking that is prudent and well informed. The primary role of the risk management function in a bank is to ensure that the total risk taken across the enterprise is no greater than the bank's ability to absorb worst-case losses within some specified confidence interval.

In its pure form, capital represents the difference between the market value of a bank's assets and the market value of its liabilities. Because capital can be viewed as a buffer against insolvency, capital adequacy is a measure of a bank's ability to remain a *going concern* under adverse conditions.

In contrast to a typical corporation, the key role of capital in a financial institution such as a bank is not primarily one of providing a source of funding for the organisation. Banks usually have ready access to funding through their deposit-taking activities, which can be increased fairly fluidly. Instead, the primary role of capital in a bank, apart from the transfer of ownership, is to act as a buffer to:

- absorb large unexpected losses;
- protect depositors and other claim holders;
- provide enough confidence to external investors and rating agencies on the financial health and viability of the firm.

A firm's credit rating can be seen as a measure of its capital adequacy and is generally linked to a specific probability that the firm will enter into default over some period of time. If we make the assumption that liabilities are riskless, then the credit rating of a firm becomes a function of the overall riskiness of its assets and the amount of capital that the bank holds. Firms which hold more capital are able to take on riskier assets than firms of similar credit rating which hold less capital.

Typically, the sources of risk within the assets of a firm are classified as follows:

- credit risk – losses associated with the default (or credit downgrade) of an obligor (a counterparty, borrower or debt issuer);
- market risk – losses associated with changes in market values;
- operational risk – losses associated with operating failures.

Capital represents an ideal metric for aggregating risks across both different asset classes and across different risk types.

III.0.1.2 Types of Capital

We can broadly classify capital into three types:²

- *Economic capital* (EC) – an estimate of the level of capital that a firm requires to operate its business with a desired target solvency level. Sometimes this is also referred to as *risk capital*.³
- *Regulatory capital* (RC) – the capital that a bank is required to hold by regulators in order to operate; this is an accounting measure defined by the regulatory authorities to act as a proxy for economic capital.
- *Book capital* (BC) – the actual physical capital held. While in its strictest definition this should be simply *equity capital*, more generally this might also include other assets like liquid debt or hybrid instruments

In practice, many firms hold book capital in excess of the required economic and even regulatory capital. This reflects both historical and practical business reasons (for example, given their size, they might be too slow to invest it effectively), as well as their more conservative view on the applicability of the models. The combined forces of deregulation and the increased market volatility in the late 1970s motivated many banks to aggressively grow market share and to acquire increasingly riskier assets on their balance sheets. This emphasis on growth precipitated a decline of capital levels throughout the 1980s that led to fears of increasing instability in the international banking system. These concerns motivated the push for the creation of international capital adequacy standards such as those ultimately established by the Basel Committee on Banking Supervision (BCBS). The imposition of the Basel I Accord in 1988 proved to be successful in its objective of increasing worldwide capital levels to desired levels by 1993 and, ultimately, to reduce the overall riskiness of the international banking system.

In general, EC is meant to reflect the true ‘fair market’ value differential between assets and liabilities, and thus it is limited by the ability to mark to market a balance sheet in a manner that is indisputable for all key constituencies – the financial institution, the regulators and the investors themselves. As such, the determination of EC has traditionally been highly institution-specific.

The foremost objective of regulations, however, is to define an unarguable standard for capital comparison that creates a level playing field across all financial institutions. Thus, regulatory capital has traditionally been defined with respect to accounting book value measures rather than

² This is a general classification, and there are various alternative definitions of capital and terminology used to describe them.

³ Some authors have used alternative definitions: for example, Matten (2000, pp. 222–223) defines economic capital as risk capital plus goodwill; Perold (2001) defines risk capital in terms of insurance (explained in Section III.0.2.6).

to market value measures (notwithstanding the fact that, in some cases, accounting practice allows balance sheet items to be reported on a market value basis).

To capture the discrepancy between fair values and market values, regulatory capital measures incorporate *ad hoc* approaches to normalise asset book values to reflect differences in risk. As such, it is recognised that regulatory capital calculations tend to contain a number of inconsistencies, which have led regulators to set prescribed levels on a conservative basis. In some cases, these inconsistencies have led to the notion of *regulatory arbitrage*, whereby investment is determined not on the basis of risk–reward optimisation but on the basis of regulatory capital–reward optimisation.

III.0.1.3 Capital as a Management Tool

Capital can be used as a powerful business management tool, since it provides a consistent metric to determine:

- risk aggregation;
- performance measurement;
- asset and business allocation.

The objective of risk-adjusted performance measurement (RAPM) is to define a consistent metric that spans all asset and risk classes, thereby providing an ‘apples to apples’ benchmark for evaluating the performance of alternative business opportunities. RAPM thus becomes an ideal tool for capital allocation purposes. By allocating the appropriate amount of EC to each asset, net expected payoffs can then be expressed as returns on capital. Each asset can, therefore, be assessed on a consistent basis, with returns adjusted appropriately in the context of the amount of risk taken on. (This is further discussed in Sections III.0.4 and III.0.5).

Risk aggregation generally refers to the development of quantitative risk measures that incorporate multiple sources of risk. The most common approach is to estimate the EC that is necessary to absorb potential losses associated with each of the risks. EC can be seen as a common measure that can be used to summarise and compare the different risks incurred by a firm, across

- different businesses and activities;
- different types of risk – market risk, credit risk and operational risk.

According to a recent study (BCBS, 2003b), the application of risk aggregation and EC methods is still in the early stages of its evolution. While some firms remain sceptical of the value of reducing all risks to a single number, many now believe that there is a need for a common metric

that allows risk–return comparisons to be made systematically across business activities whose mix of risks may be quite different (e.g., insurance versus trading). However, there remains a wide variation in the manner in which aggregated risk measures such as EC are used for risk management decision making in practice today.

III.0.2 Economic Capital

Economic capital acts as a buffer that provides protection against all the credit, market, operational and business risks faced by an institution. EC is set at a confidence level that is less than 100% (e.g., 99.9%), since it would be too costly to operate at the 100% level. The confidence interval is chosen as a trade-off *between* providing high returns on capital for shareholders *and* providing protection to the debt holders (and achieving a desired rating) as well as confidence to other claim holders, such as depositors.

In so far as EC reflects the amount of capital required to maintain a firm’s target capital rating, the confidence interval can be defined at a very high quantile of the loss distribution. For example, to achieve a target S&P credit rating of BB, the probability of default over the next year for the firm cannot be greater than 3.0%, so the quantile should be set at least at 97%. In contrast, for the same firm to achieve a target S&P credit rating of BBB, it must lower its probability of default to be at most 0.5%, corresponding to the 99.5% quantile of the loss distribution. Given the desire to achieve a BBB rating and to remain solvent 99.5% of the time, a firm must have enough capital to sustain a ‘0.5% worst-case loss’ over a one-year time horizon; that is, 99.5% of the time the future value of the non-defaulted assets must be at least equal to the future value of the liabilities. Example III.0.1 below gives a simple outline of how this could be achieved.

III.0.2.1 Understanding Economic Capital

Denote by A_t and D_t the market values (at time t) of the assets and liabilities, respectively. The available capital C_t for the current time, $t = 0$, and at the end of one year, $t = 1$, can be expressed as

$$\begin{aligned} C_0 &= A_0 - D_0, \\ C_1 &= A_1 - D_1. \end{aligned} \tag{III.0.1}$$

If the nominal returns on the assets and liabilities are equal to r_A and r_D , respectively, then a worst-case loss from all sources, l (i.e., when $C_1 = 0$ for a given confidence interval), would result in the value of assets at $t = 1$ just being sufficient to cover the value of debt in $t = 1$. Then

$$C_1 = 0 = A_0 (1 + r_A)(1 - l) - D_0(1 + r_D).$$

Thus the maximum amount of debt allowable to sustain solvency under the worst-case scenario cannot exceed

$$D_0 = A_0 (1 + r_A)(1 - l)/(1 + r_D). \quad (\text{III.0.2})$$

Since EC_0 is the minimum amount of capital required to sustain such a loss, it is given by:

$$EC_0 = A_0 (1 - [(1 + r_A)(1 - l)/(1 + r_D)]). \quad (\text{III.0.3})$$

Hence the minimum amount of EC a financial institution must take on in order to avoid insolvency increases as the level of the worse-case loss l increases.

The *expected return on EC* over the period from $t = 0$ to $t = 1$ is given by

$$[E(EC_1)/EC_0] - 1.$$

The expected return on EC reflects the impact of leverage on risk and reward. An increase in expected returns (to compensate for increased risk) is reflected in the numerator, while the increase in risk is reflected in the denominator (the current EC).

Equation (III.0.3) is often expressed in terms of *value-at-risk* notation in the following manner:

$$EC_0 = A_0 - VaR/(1+r_D), \quad (\text{III.0.4})$$

where VaR represents the A_i value associated with the worst-case loss, l , corresponding to the appropriate ($x\%$) confidence interval.

For ease of presentation, consider the case where credit risk is the sole source of business risk to which the firm is exposed. Returning to equation (III.0.3), under the simplifying assumption that the spread between the nominal return on the assets and the return on the liabilities is roughly equal to the expected default loss, u , then,

$$\begin{aligned} EC_0 &= A_0 \{1 - (1 + r_D)(1 + u)(1 - l)/(1 + r_D)\} \\ &= A_0 \{1 - (1 + u)(1 - l)\}. \end{aligned} \quad (\text{III.0.5})$$

By then ignoring second-order effects, equation (III.0.5) simplifies to the following more familiar expression for economic capital:

$$EC_0 \approx A_0 \cdot (l - u). \quad (\text{III.0.6})$$

This relationship is illustrated with respect to a default loss distribution in Figure III.0.1.

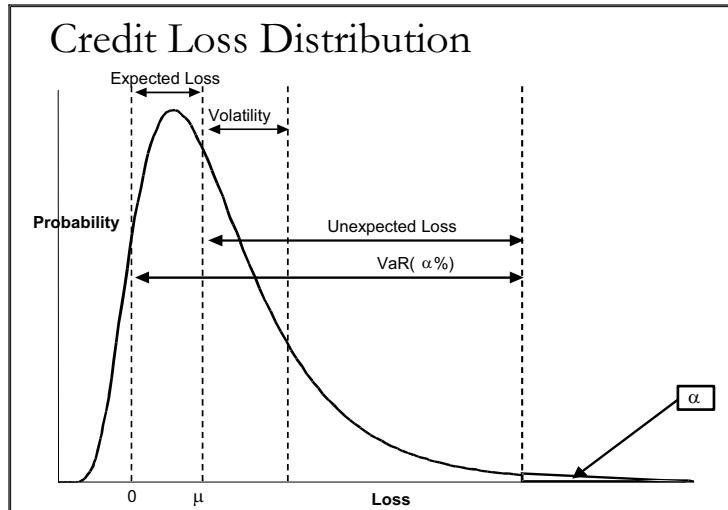


Figure III.0.1: Credit loss distribution: expected and unexpected losses

Expressions (III.0.4) and (III.0.6) highlight the link between VaR measures and EC. The simplifying assumption leading to equation (III.0.6) and illustrated in Figure III.0.1 is the approach commonly taken by practitioners and generally leads to conservative estimates (for a detailed discussion, see Kupiec, 2002). Thus, in its most common definition, EC is defined to absorb only *unexpected losses* (UL) up to a certain confidence level (i.e., $A_0(l - u)$). *Credit reserves* are traditionally set aside to absorb *expected losses* (EL) over the period (i.e., A_0u). More precisely, equation (III.0.4) shows that the VaR measure appropriate for EC should in fact measure losses relative to the assets' initial mark-to-market (MtM) value and *not* relative to the EL in its end-of-period distribution. Also, the VaR measure should explicitly account for the interest payments on the funding debt. While the UL approximation has very little effect on market risk, where the horizon is short (and EL is small) it may have a higher impact in credit risk.

Example III.0.1

Consider a BBB-rated firm (or a firm that has targeted a BBB rating). Suppose the firm has liabilities consisting of $D_0 = \$92$ million in deposits, with a cost of debt of $r_D = 5\%$, which have been invested in $A_0 = \$100$ million of assets (40% at a nominal return of 6.75% and 60% at a nominal return of 7%). The weighted average nominal return across the \$100 million in total assets is $r_A = 6.9\%$, representing a compounded spread of 1.81%. If the nominal values of the assets and liabilities are equal to the market values, then the current capital for this firm is calculated as $C_0 = \$8$ million (the difference between the market value of the assets and the market value of the liabilities).

Assume that, in a ‘0.5% worse-case scenario’ the firm has a potential for a loss of 15% in the value of total assets. Under this scenario, the firm will become insolvent as the value of the assets will be $A_1 = \$100 \text{ million} \times 1.069 \times 0.85 = \90.9 million , while the value of the liabilities will be $D_1 = \$92 \text{ million} \times 1.05 = \96.6 million , giving a capital shortfall of \$5.7 million.

From equation (III.0.3), the minimum amount of capital the firm must hold to avoid insolvency in the worst-case scenario is

$$EC_0 = A_0 (1 - [(1 + r_A)(1 - l)/(1 + r_D)]) = 100(1 - [1.069(1 - 0.15)/1.05]) = 13.46.$$

Therefore, for the firm to improve its capital adequacy to the desired level, it must increase its capital from \$8 million to \$13.46 million. The shareholders should be 99.5% sure that such an increase in capital will ensure solvency from $t = 0$ to $t = 1$.

III.0.2.2 The Top-Down Approach to Calculating Economic Capital

EC can be seen as a common measure that can be used to summarise and compare the different risks incurred by a firm, across different businesses and activities, and across different types of risk: market, credit and operational risk. At the enterprise level, EC can be estimated based on aggregate information of the firm’s performance. Such ‘top-down approaches’ generally use one of two types of information: earnings or stock prices.

III.0.2.2.1 Top-Down Earnings Volatility Approach

A top-down approach based on a firm’s earnings makes the simplifying assumption that the market value of capital is equal to the value of a perpetual stream of expected earnings. In other words, by assuming that all expected future earnings of the firm are equal to the next period’s expected earnings, the value of capital can be expressed as

$$C_0 = \text{Expected earnings} / k,$$

where k represents the required return associated with the riskiness of the earnings. As the determination of EC is based on the ability to sustain a worst-case loss associated with a given confidence interval,

$$EC_0 = EaR / k,$$

where EaR represents the difference between expected earnings and the earnings under the worst-case scenario, for a given confidence interval (see Saita, 2003). Often this approach relies on the additional assumption that earnings are normally distributed, and thus the confidence interval can be determined as a multiple of the standard deviation.

Limitations of the earnings volatility approach include the following:

- It requires historical performance data for reliable estimates of the mean and standard deviation of earnings; few companies have enough data to yield reliable estimates.
- It does not link EC directly to the sources of risk.
- In general, it does not naturally allow capital to be separated out into its market, credit and operational risk components, nor across different business lines or activities.

III.0.2.2.2 Top-Down Option-Theoretic Approach

A top-down approach based on the Black–Scholes–Merton (BSM) framework (see Chapter III.B.5) assumes that the market value of capital can be modelled as a call option on the value of the firm's assets where the strike price is the notional value of the debt. If the value of assets at the end of the period ($t = 1$) is greater than the value of the debt, then the value of capital is equal to the difference between the value of the assets and the debt; otherwise (in the case of insolvency), it is equal to zero. Using this approach assumes we have the following information available:

- the current market value and volatility of the company's net assets;
- the time horizon (e.g., the average duration of the firm's assets);
- the risk-free interest rate (maturity corresponding to the time horizon);
- the default threshold (the asset level at which the debt holders demand repayment and bankruptcy can occur).

The BSM model allows us to estimate the *implied* probability of insolvency for the firm over the period from $t = 0$ to $t = 1$. The EC can then be determined on the basis of reapplying the BSM model for a level of debt that ensures, even under the worst-case scenario (at a given confidence interval), that the firm remains solvent.

An advantage of this approach over the one based on EaR is the availability of stock market data. However, several simplifications regarding the capital structure and model assumptions must be made to apply this tool in practice. Similar to the EaR approach, a key limitation is that it does not allow the separation of capital into different risks such as market, credit and operational risks, nor does it suggest how to allocate it across different business lines or activities.

III.0.2.3 The Bottom-Up Approach to Calculating Economic Capital

In this approach, EC is estimated by modelling individual transactions and businesses and then aggregating the risks using advanced statistical portfolio models and stress testing. The bottom-up approach has now become best practice and, in contrast to the top-down approaches, provides greater transparency with regard to isolating credit risk, market risk and operational risk

capital. Furthermore, it naturally accommodates various methodologies to allocate capital to individual businesses, activities and transactions.

In a bottom-up approach, the estimation of enterprise EC requires consolidation of risks at two levels:

- First, it computes market risk, credit risk and operational risk at the enterprise level. To achieve this, a firm might use an internal VaR model for market risk, a credit VaR methodology for credit risk and a loss-distribution approach for operational risk.
- Then, at the second level, the firm must consolidate the capital across these risks.

To estimate total capital, it is currently common practice to add up the credit risk capital, market risk capital and operational risk capital. This produces a conservative capital measure (basically assuming that the risks are perfectly positively correlated). However, today, many firms are now devoting considerable effort to measuring the correlations between these risks (and hence the levels of diversification), as well as developing frameworks to measure these risks in a more integrated way.

III.0.2.4 Stress Testing of Portfolio Losses and Economic Capital

In addition to the statistical approaches inherent in credit portfolio models, practitioners usually use stress testing as an important part of their EC methodology (see Chapter III.A.4). Commonly, the stress-testing methodology involves the development of one or several specific adverse scenarios, which are judged to be extreme (falling beyond the desired confidence level). Current portfolio losses are then assessed against these specific scenarios. Stress scenarios may be based on historical experience or management judgment.

The translation of the specific stress scenario losses, and the combination of stress testing and statistical measures, to develop EC measures is today more an art than a science, largely based on management's objectives and judgement. In essence, firms make their own decision on the relative weights of the statistical and stress test results in estimating the amount of EC required to support a portfolio. For example, an institution might assign the EC for market risk as 50% times the 99% VaR plus 50% the loss outcome from some stress scenarios (thus normally being higher than the actual 99% VaR).

III.0.2.5 Enterprise Capital Practices – Aggregation

A large firm, such as a bank, acquires different types of financial risk through various businesses and activities. Capital is indeed a powerful tool for understanding, comparing and aggregating

different types of risk to determine the overall health of the firm and to support better business decisions.

Such an institution is likely to have separate methodologies to measure market risk, credit risk and operational risk. In addition, it is likely that the institution has different methodologies to measure the credit risk of its larger commercial loans or its retail credits. More generally, a firm may have any number of methodologies for various risks and segments. If each type of risk is modelled separately, then the amounts of EC estimated for each need to be combined to obtain an enterprise capital amount. In making the combination, the firm needs to incorporate, either implicitly or explicitly, various correlation assumptions. The methods of aggregation are as follows:

- Sum of stand-alone capital for each business unit and type of risk. This methodology essentially assumes perfect correlation across business lines and risk types and does not allow for diversification from them.
- *Ad hoc* or top-down estimates of cross-business and cross-risk correlation. In order to allow for some cross-business and cross-risk diversification, a firm might aggregate the individual stand-alone capital estimates using analytical models and simple cross-business (asset) correlation estimates.

The enterprise aggregation of capital is still in its infancy and is a topic of much research today.

III.0.2.6 Economic Capital as Insurance for the Value of the Firm

Standard practice is to define EC as a buffer to cover unexpected losses. Thus it is defined in terms of the tail of the loss distribution, using measures such as VaR. Alternatively, some economists have used the term ‘risk capital’ to define capital in economic terms (Merton and Perold, 1993; Perold, 2001): risk capital is the smallest amount that can be invested to insure the value of the firm’s net assets⁴ against loss in value relative to a risk-free investment.

For example, under this definition, the risk capital of a long US Treasury bond position is the value of a put option with strike equal to the forward price of the bond. As pointed out by Perold (2001), in general, the put option accounts for the full distribution of losses, whereas VaR ignores the magnitude of outcomes conditional on being in the extreme tail of the distribution. When returns are normally distributed, the value of such a put option is approximately proportional to the standard deviation of the return on the bond, and thus is approximately proportional to VaR.

⁴ ‘Net assets’ refers to ‘gross assets minus customer liabilities (swaps, insurance contracts, etc.), valued as if these liabilities are default-free.

III.0.3 Regulatory Capital

This section considers the key concepts and objectives behind regulatory capital, as well as the main principles used in regulatory capital calculations in the Basel I Accord and the latest proposals of the current Basel II Accord.

III.0.3.1 Regulatory Capital Principles

In this subsection, we focus mainly on the capital regulation in the banking industry. As defined in Section III.0.1, regulatory capital refers to the capital that an institution is required to hold by regulators in order to operate. It is largely an accounting measure defined by the regulatory authorities to act as a proxy for economic capital. Capital adequacy is generally the single most important financial measure used by banking supervisors when examining the financial soundness of an institution.

As also mentioned in Section III.0.1, from an *internal* bank perspective, capital is designed as a buffer to absorb large unexpected losses, protect depositors and other claim holders, and provide enough confidence to external investors and rating agencies on the financial health and viability of the firm. In contrast, from the *external* perspective of the regulator, capital adequacy requirements fulfil two objectives:

- *Reducing systemic risk:* to safeguard the security of the banking system and ensure its ongoing viability. In a sense, national governments act as guarantors. They have an interest in ensuring that banks remain capable of meeting their obligations and in minimising potential systemic effects on the economy. Regulatory capital helps to ensure that banks bear their share of the burden, otherwise borne by national governments.
- *Creating a level playing field:* to ensure a more even playing field for internationally active banks, by submitting all banks to (roughly) the same rules.

As we review the key concepts in regulatory capital, it is important to highlight two points:

- Regulatory requirements are continuously changing, and it is vital for practitioners to be familiar with both the latest regulations and the specific requirements in each jurisdiction.
- While the general intention is to make regulatory capital more risk-sensitive and align it more closely to economic capital, it is important to understand the limitations of using it directly for managing risk, measuring performance and pricing credits.

The overall objective should be to set up an enterprise risk management framework, which measures economic capital and reconciles it with regulatory capital.

III.0.3.2 The Basel Committee of Banking Supervision and the Basel Accord

A key cornerstone of international banking capital regulation is the Basel Committee on Banking Supervision, which first introduced the framework for international capital adequacy standards. This framework has been adopted as the underlying structure of all bank capital adequacy regulations throughout the G10, as well as many other countries around the world. Today, over 100 countries are expected to implement the latest guidelines set by the BCBS, also referred to as the Basel II accord).

Summary Chronology of Banking Regulatory Capital

- 1988 – The BCBS introduces the framework for international capital adequacy standards. It is adopted throughout the G10, as well as in over 100 other countries (BCBS, 1988). Commonly referred to as the *Basel I Accord* or the *BIS I Accord*, it was the first step in establishing a level playing field across member countries for internationally active banks. The 1988 accord focused mainly on credit risk.
- 1995 – An amendment to the initial accord further allows banks to reduce ‘credit-equivalent exposures’ when netting agreements are in place (BCBS, 1995).
- 1996 – The 1996 amendment⁵ extends the capital requirements to include risk-based capital for the market risk in the trading book (BCBS, 1996).
- 1999 – The BCBS issues a proposal for a new capital adequacy framework to replace the 1988 Basel I Accord. This is commonly referred to as the *Basel II Accord* or *BIS II Accord*. The new accord attempts to improve the capital adequacy framework by substantially increasing the risk sensitivity of the minimum capital requirements, and also encompassing a supervisory review and market discipline principles. Under the proposal, banks are required specifically to allocate capital against operational risks for the first time.
- 2000–2003 – The BCBS releases various consultation documents and conducts major data collection exercises called quantitative impact studies (QIS), intended to gather information to assess whether it has met its goals.
- April 2003 – The BCBS releases the third consultative paper (CP3) on the new Basel Accord (BCBS, 2003a).

⁵ Sometimes referred to as *BIS 98*, after its date of implementation.

- June 2004 – The final version of the Basel II Accord is published (BCBS, 2004).
- 2006–2007 – Currently scheduled implementation of Basel II.

All the papers from the BCBS can be downloaded from www.bis.org.

III.0.3.3 Basel I Regulation

The 1988 accord focused mainly on credit risk, establishing minimum capital standards that linked capital requirements to the credit exposures of banks. Prior to its implementation in 1992, bank capital was regulated through simple, *ad hoc* capital standards. While generally prescriptive, Basel I left various choices to be made by local regulators, thus resulting in several variations of the implementation across jurisdictions. The 1996 amendment further extended the capital requirements to include risk-based capital for the market risk in the trading book. Basel I does not cover capital charges for operational risk.

III.0.3.3.1 Minimum Capital Requirements under Basel I

Capital requirements under Basel I are the sum of:

- credit risk capital charge, which applies to all positions in the trading and banking books (including OTC derivatives and balance sheet commitments);
- market risk capital charge for the trading book portfolio and off-balance sheet items.

For market risk capital, the accord allows, in addition to a standardised method, the use of internal VaR models covering both general market risk (or systemic risk) and specific risk. Specific VaR applies to both equities and bonds. For bonds it covers the risk of defaults, migration and changes in spreads. The reader is referred to Chapter III.A.2 for the basics of market risk VaR.

The regulatory charge for banks using internal market risk models is given by

$$\text{Market Risk Capital} = [M_{MR} \cdot VaR + M_{SR} \cdot \text{Specific VaR}] \cdot \frac{\text{Trigger}}{8}, \quad (\text{III.0.7})$$

where *VaR* and *Specific VaR* denote, respectively, the 99% market VaR and specific VaR over a 10-day horizon, and M_{MR} and M_{SR} are multipliers designed to adjust the capital to cover for modelling errors and reward the quality of the models. The first one ranges between 3 and 4, and the second one between 4 and 5. Finally, the *Trigger* is related to quality of controls in the bank. Currently it is set to 8 in North America and between 8 and 25 in the UK.

The methodology for credit capital is simple. Minimum capital requirements are obtained by multiplying the sum of all the risk-weighted assets by the capital adequacy ratio of 8% (also referred to as the *Cook ratio*):

$$\text{Capital} = \left(\sum_k \text{RWA}_k \right) \cdot 8\% . \quad (\text{III.0.8})$$

Thus the calculation of credit regulatory requirements has three steps: converting exposures to credit equivalent assets; computing loan equivalents for off-balance sheet and OTC portfolios; and applying the capital adequacy ratio. This is described in greater detail in Chapter III.B.6.

A great strength of Basel I is the simplicity of the framework. This has allowed it to be implemented in countries with different banking and accounting practices. Thus, it has been quite successful in achieving its two general objectives (to safeguard the stability of the banking system and to ensure an level playing field internationally).

Its simplicity also has been its major weakness, as the accord does not effectively align regulatory capital requirements closely with an institution's risk. For example, some criticisms on the credit risk capital include the lack of proper differentiation for credit quality and maturity, insufficient incentives for credit mitigations techniques, and lack of recognition of portfolio effects (these are discussed briefly in Chapter III.B.6).

III.0.3.3.2 Regulatory Arbitrage under Basel I

The lack of differentiation in the accord, together with the financial engineering advances in credit risk over the last decade, have lead to the development of a *regulatory capital arbitrage* industry. This refers to the process by which regulatory capital is reduced through instruments such as credit derivatives or securitisation, without an equivalent reduction of the actual risk being taken. Through regulatory arbitrage instruments, for example, banks typically transfer low-risk exposures from their banking book to their trading book, or simply place them outside the regulated banking system.

III.0.3.3.3 Meeting Capital Adequacy Requirements

Available regulatory credit capital is divided into two categories:

- Tier 1 capital: essentially shareholder funds – equity – and retained earnings.
- Tier 2 capital: long-term subordinated debt, other qualifying hybrid instruments and reserves (such as loan loss reserves).

From a regulatory perspective, Tier 1 capital must cover at least 50% of the total capital; that is, Tier 2 cannot exceed Tier 1 capital. In addition, the subordinated debt included in Tier 2 cannot exceed 50% of the Tier 1 capital.⁶

Capital adequacy is generally expressed as a ratio. For example, an 8% capital ratio means that the total Tier 1 and Tier 2 capital is 8% of the risk-weighted assets (RWA). A 6% Tier 1 capital ratio refers to Tier 1 capital being 6% of the RWA.

III.0.3.4 Basel II Accord – Latest Proposals

The final version of the Basel II Accord was published in June 2004 (BCBS, 2004).⁷ Its implementation will take effect between the end of 2006 and the end of 2007.

In this subsection we present a brief summary of the basic principles of Basel II; the key formulae for minimum capital requirements for credit risk are given in Chapter III.B.6. For greater detail, the reader is referred to the BCBS papers.

Basel II attempts to improve capital adequacy framework along two important dimensions:

- First, the development of a capital regulation that encompasses not only minimum capital requirements, but also supervisory review and market discipline.
- Second, a substantial increase in the risk sensitivity of the minimum capital requirements.

The new accord intends to foster a strong emphasis on risk management and to encourage ongoing improvements in banks' risk assessment capabilities. This is to be accomplished by closely aligning banks' capital requirements with prevailing modern risk management practices, and by ensuring that this emphasis on risk makes its way into supervisory practices and into market discipline through enhanced risk- and capital-related disclosures.

The Basel II Accord consists of three pillars: *minimum capital requirements*, *supervisory review*, and *market discipline*. We briefly summarise these below and then present the key principles behind the computation of minimum capital requirements.

III.0.3.4.1 Pillar 1 - Minimum Capital Requirements

Minimum capital requirements consist of three components:

1. definition of capital (no major changes from the 1988 accord);

⁶ A Tier 3 capital was introduced with the market risk requirements. Short-term subordinated debt can be used to meet market risk requirements as well, but not credit risk.

⁷ A small number of open issues are still to be resolved during 2004.

2. definition of RWA;
3. minimum ratio of capital/RWA (remains 8%).

Basel II proposes to modify the definition of risk-weighted assets in two areas:

- substantive changes to the treatment of credit risk relative to the Basel I Accord;
- the introduction of an explicit treatment of operational risk that will result in a measure of operational risk being included in the denominator of a bank's capital ratio.

Basel II moves away from a one-size-fits-all approach to the measurement of risk, through the introduction of three distinct options for the calculation of credit risk and three others for operational risk. These approaches present increasing complexity and risk-sensitivity. Banks and supervisors can thus select the approaches that are most appropriate to the stage of development of banks' operations and of the financial market infrastructure. Chapter III.B.6 briefly reviews the three credit risk approaches. The operational risk approaches can be found in Chapter III.C.3.

III.0.3.4.2 Pillar 2 - Supervisory Review

The second pillar is based on a series of guiding principles, which point to the need for banks to assess their capital adequacy positions relative to their overall risks, and for supervisors to review and take appropriate actions in response to those assessments. Banks under internal ratings-based credit models will be required to demonstrate that they use the outputs of those models not only for minimum capital requirements but also to manage their business. The inclusion of supervisory review provides benefits through its emphasis on strong risk assessment capabilities by banks and supervisors alike. Important new components of Pillar II also include the treatment of stress testing, concentration risk and the residual risks arising from the use of collateral, guarantees and credit derivatives as well as specific securitisation exposures (these are discussed further in Chapter III.B.6).

III.0.3.4.3 Pillar 3 - Market Discipline

Also referred to as *public disclosure*, the third pillar aims to encourage safe and sound banking practices through effective market disclosures of capital levels and risk exposures. This will help market participants assess better a bank's ability to remain solvent.

III.0.3.5 A Simple Derivation of Regulatory Capital

Recognising that the simple difference between the value of assets and the value of liabilities in accounting value terms is not a good indicator of the true difference in market value terms has led regulators to make appropriate adjustments in the calculation of regulatory capital. In this section, we follow a similar approach to Section III.0.2.1 to understand regulatory capital.

For a typical bank balance sheet, the difference between assets and liabilities includes general provisions GP , and reserves R_0 , as well as the book value of equity $E_0\{BV\}$:

$$A_0\{B/S\} - L_0\{BV\} = GP_0 + E_0\{BV\} + R_0. \quad (\text{III.0.9})$$

The amount of available capital for regulatory purposes, however, can be defined loosely as the difference between *total* assets (balance sheet as well as non-balance sheet) and only that component of total liabilities where non-payment of returns defines insolvency. The amount of available capital can thus be represented as follows:

$$RC_0 = A_0\{B/S\} + A_0\{non-B/S\} - D_0\{BV\}. \quad (\text{III.0.10})$$

Balance sheet assets are the net of the book value of assets and any special provisions to account for defaulted or nearly-defaulted positions, while non-balance sheet assets consist of any revaluations, RV_0 , to market value as well as any undisclosed profits, UP_0 . That is,

$$\begin{aligned} A_0\{B/S\} &= A_0\{BV\} - SP_0 \\ A_0\{non-B/S\} &= RV_0 + UP_0. \end{aligned} \quad (\text{III.0.11})$$

Those liabilities whose non-payment constitutes insolvency represent the difference between total liabilities and the quasi-debt, QD_0 , that combines both debt and equity features:⁸

$$D_0\{BV\} = L_0\{BV\} - QD_0\{BV\}. \quad (\text{III.0.12})$$

Combining these relationships provides a very straightforward definition of regulatory capital in book value terms that is designed to be as good a proxy as possible for true market valuation:

$$RC_0 = E_0\{BV\} + R_0 + QD_0\{BV\} + GP_0. \quad (\text{III.0.13})$$

The first two terms of this definition of regulatory capital are referred to as Tier 1 capital, while the second two terms are referred to as Tier 2 capital. Capital adequacy standards are based on minimum requirements for each of the two tiers of capital.

Note that the above adjustments to book value measures still do not adequately capture the true market values and, hence, the true riskiness of the assets. Therefore, standards on regulatory capital are typically expressed as minimum percentages of *risk-weighted* assets rather than total assets. An RWA is expressed as a percentage of the nominal value of a balance sheet asset. For example, under the first Basel Accord loans to private companies are assigned a risk weight of 100%, while loans to banks in OECD countries are assigned a risk weight of 20%, reflecting the perceived differences in risk between the two categories of borrowers.

⁸ Non-payment of quasi-debt does not imply insolvency, at least for a period of time.

III.0.4 Capital Allocation and Risk Contributions

We discuss the importance for allocating economic capital to different business units in a firm – or to the constituents of a ‘portfolio’ – and the key methodologies to compute contributions to EC.

III.0.4.1 Capital Allocation

In addition to computing the total EC for a firm or portfolio, it is important to develop general methodologies to

- attribute this capital *a posteriori* to various ‘sub-portfolios’, such as the firm’s activities, business units and even individual transactions, and
- allocate it *a priori* in an optimal fashion, to maximise risk-adjusted returns.

EC allocation down to the portfolio is required for:

- management decision support and business planning;
- performance measurement and risk-based compensation;
- pricing, profitability assessment and limits;
- building optimal risk–return portfolios and strategies.

From a strategic management perspective, the allocation of EC to business units, activities and transactions is an issue that is receiving significant attention in the industry today. There are two views prevalent among firms:

- Diversification benefits should not be passed down to the business units. Rather, each unit is expected to operate on a stand-alone basis.
- An ‘optimal’ level of group risk taking can be achieved only when diversification benefits are allocated to at least the major business units (and perhaps even to the transaction level). Thus, it is preferable for each business unit to be assigned an EC allocation closer to its ‘marginal contribution’ to the total EC.

In the general case, the sum of the stand-alone EC for each asset or business does not equal the total portfolio EC. Indeed, it is higher, since there are diversification benefits. Thus, it is important to devise a general methodology to assign capital to individual business units, activities and assets, which explicitly allocates the diversification benefits of the portfolio.

III.0.4.2 Risk Contribution Methodologies for EC Allocation

There is no unique method to allocate EC down a portfolio, and thus it is important to understand how risk contribution tools can be applied to EC allocation decisions. Whether it is

EC contributions to business units, arbitrary sub-portfolios or assets, we can classify the allocation methodologies which are currently used in practice into three categories: *stand-alone* EC contributions; *incremental* EC contributions; and *marginal* EC contributions.⁹ Every methodology has its advantages and disadvantages, and might be more appropriate for a particular managerial application.

III.0.4.2.1 Stand-alone EC Contributions

An individual business or sub-portfolio is assigned the amount of capital that it would consume on a stand-alone basis (e.g., if it were an independent firm). As such, it does not reflect the beneficial effect of diversification. The resulting sum of stand-alone capital for the individual business units, activities or sub-portfolios is generally greater than the total EC for the firm.

III.0.4.2.2 Incremental EC Contributions

This method is also referred sometimes as the *discrete marginal EC allocation* method. Under this method the EC allocated to a business unit or sub-portfolio attempts to capture an appropriate amount of risk capital that the unit contributes to the entire firm's capital requirements. It is calculated by taking the EC computed for the entire firm (including the business unit or sub-portfolio) and subtracting from it EC for the firm without the business unit or sub-portfolio. This methodology thus captures exactly the amount of capital that would be released if the business unit were sold or added (everything else remaining the same).

Incremental EC is a natural measure for evaluating the risk of acquisitions or divestitures.¹⁰ But, while very intuitive, a disadvantage of this methodology is that it is not *additive*. While it does capture the benefits of diversification, the sum of incremental EC for all the firm's business units (activities or sub-portfolios) is smaller than (or equal to) total EC for the firm.

III.0.4.2.3 Marginal EC Contributions

It would be useful to obtain measures of risk contributions that are additive. Sometimes referred to as *diversified EC contributions*, such measures are intended to capture the amount of the firm's total capital that should be allocated to a particular business or sub-portfolio when viewed as part of a multi-business firm. They are specifically designed to allocate the diversification benefit among the business units and activities, in the form of reduced EC. Thus, by construction, the

⁹ The reader is cautioned that there is currently no universal terminology for these methodologies in the literature. As defined here, *incremental capital (risk) contributions* are sometimes also referred to as *marginal capital (risk) contributions* or *discrete marginal capital (risk) contributions*. *Marginal risk contributions*, as termed here, are sometimes also referred to as *diversified capital (risk) contributions* or, more precisely, *continuous marginal capital (risk) contributions* (Smithson, 2003).

¹⁰ See, for example, Perold (2001) – note that the author refers to this as 'marginal' EC.

sum of diversified EC for all the firm’s business units and activities is equal to total EC for the firm.

There are various methodologies that produce additive risk contributions. The most widespread and, perhaps, practical methodology is the one based on *marginal risk contributions*. While most explanations of this risk decomposition methodology are based on the use of volatility (or standard deviation) as a risk measure, the methodology is quite general and applicable to other risk measures. Volatility-based contributions are common practice today (see Smithson, 2003). However, such allocations can be ineffective for credit and operational risk, given the non-normality of their loss distributions. Industry best practices are shifting towards allocations based on VaR or expected shortfall (ES) – see Chapters III.A.2, III.A.3 and III.B.5.¹¹

An additive decomposition of EC is of the form:

$$EC = \sum_i EC_i , \quad (III.0.14)$$

where EC_i denotes the EC contribution of business unit or sub-portfolio i . We can then define the percentage risk contribution of the i th business unit or sub-portfolio as:

$$EC\ Contrib_i = \frac{EC_i}{EC} \cdot 100\% . \quad (III.0.15)$$

Denoting by x_i the size of the i th business unit or sub-portfolio, one can show that for EC based on volatility, VaR or ES:¹²

$$EC_i = \frac{\partial EC(x)}{\partial x_i} \cdot x_i . \quad (III.0.16)$$

That is, an EC marginal contribution is the product of the size of business unit i and the rate of change of EC with respect to that position. This product essentially represents the rate of change of EC with respect to a small (marginal) percentage change in the size of the unit.

Marginal EC contributions require the computation of the first derivative of the risk measure with respect to the size of each unit. When the risk measure used is volatility, they can be computed analytically and are simply given by the covariance of losses of that business unit with the overall portfolio divided by the volatility of losses (see Praschnik *et al.*, 2001; Smithson, 2003):

¹¹ While VaR is defined as a loss which cannot be exceeded $x\%$ of the time (a quantile of the loss distribution), ES is commonly defined as the expected loss, conditional on reaching at least an $x\%$ loss (i.e., it is the average of the $x\%$ largest losses). Sometimes ES is also referred to as ‘tail conditional expectation’ or ‘conditional VaR’. For some discussions on the use of ES and VaR for capital allocation, see Kalkbrener *et al.* (2004) and Mausser and Rosen (2004).

¹² More formally, if the risk measure is homogeneous of degree 1 and differentiable, this follows from Euler’s theorem. This is a requirement of coherent risk measures (Artzner *et al.*, 1999).

$$EC\ Contrib_i = Cov(L_i, L) / \sigma(L).$$

The general theory behind the definition and computation of these derivatives in terms of quantile measures (VaR, ES) has been also developed in the last few years (see Gouriéroux *et al.*, 2000; Tasche, 2000, 2002; see also Chapter III.B.5).

It is important to stress that these contributions must be interpreted on a marginal basis. Marginal EC contributions are very general and are best suited to understand the amount of capital to be consumed by an instrument or portfolio (which really is small compared to the whole firm). They also naturally explain how to move EC from one business to another (on a marginal basis). Proponents of marginal EC approaches point out that incremental EC always under-allocates total firm EC and that, even if the incremental EC allocations were scaled up, the signals are potentially misleading. However, marginal EC is likely to be suboptimal for analysing the addition or removal on an entire business, which is not marginal to the firm.

Example III.0.2: Capital Allocation Methods

Table III.0.1 illustrates the different capital allocation methods for a simple firm consisting of three business lines. Assume, for simplicity that the total losses over one year for each business are normally distributed, and that they are uncorrelated. The stand-alone capital of each line is, respectively, \$50 million, \$30 million and \$20 million. The total stand-alone capital is thus \$100 million. The total economic capital of the firm is simply given by

$$EC = \sqrt{EC_1^2 + EC_2^2 + EC_3^2} = 61.64 \text{ million.}$$

Table III.0.1: Capital allocation methods for a simple firm

	Stand-alone Capital (m)	% Stand-alone Contributions	Incremental Capital (m)	% Incremental Capital Contributions	Marginal Capital (m)	Marginal Capital Contributions
Business 1	50.00	50.0%	25.59	69.7%	40.56	65.8%
Business 2	30.00	30.0%	7.79	21.2%	14.60	23.7%
Business 3	20.00	20.0%	3.33	9.1%	6.49	10.5%
Total	100.00	100.0%	36.72	100.0%	61.64	100.0%
% EC						
	162.2%		59.6%		100.0%	

The last line of this table gives the total capital as a percentage of EC. Notice that the total stand-alone capital represents a 62% increase of EC. Column 4 (and 5) in the table give the incremental capital for each business (in money terms and percentage contributions). The sum of incremental

capital is only 59.6% of EC. Finally, the last two columns give the marginal contributions (in money and percentage terms). Marginal contributions add up to the total EC. Note in particular that the stand-alone percentage contributions for each business differ meaningfully from the marginal contributions. While the largest business (business 1) contributes one half of the stand alone capital, it represents almost two-thirds of the EC on a marginal basis. This can be understood from the fact that as the biggest unit, increasing its share of the portfolio also marginally increases the overall risk. To diversify risk in an optimal way, one would rather increase the share of the smaller units.

III.0.4.2.4 Alternative Methods for Additive Contributions¹³

Recently there has been discussion of several alternative methods arising from game theory, which allocate the diversification benefits across the portfolio and yield additive risk contributions (see Denault, 2001; Koyluoglu and Stoker, 2002). Game-theoretic tools are commonly applied to problems involving the attribution of cost among a group and, hence, can potentially offer a useful framework for identifying ‘fair’ EC attributions. An example of these tools is the Shapley method, which describes how coalitions can be formed so that a group of units benefits more as a group than if each works separately. In this approach, the EC assigned to a unit becomes a cost, and each unit attempts to minimise its cost. A player, of course, leaves the coalition if it is attributed a larger share of EC than its own stand-alone EC. This method is computationally intensive, and may be impractical for problems with even a small number of business units. A variant called the Aumann–Shapley method further allows for ‘fractional’ units and requires less computation; thus, it is potentially more practical. Under most (but not all) conditions, both these methods may yield similar results to marginal contributions. While these methods are today receiving some academic attention, they are mostly not yet used in practice by financial institutions.

III.0.5 RAROC and Risk-Adjusted Performance

We describe the objectives of risk-adjusted performance measurement, the role of capital allocation, and the basic principles of risk-adjusted return on capital.

III.0.5.1 Objectives of RAPM

Banks traditionally measured their performance relative to their balance sheet assets, either simply with respect to overall asset size, or by bringing in the notion of profitability, with respect to

¹³ This section is added for completeness and is not mandatory.

returns on assets (ROA). There are a number of issues that make these approaches far from ideal, of which two are quite fundamental.

The first issue is that by focusing solely on assets, the performance impact of financial leverage is ignored as it pertains to managing risk and return for shareholders. In addition to the leverage effect, today many banks have off-balance sheet exposures that are ignored or, at least, not well captured by the assets as represented on a typical balance sheet.

The second fundamental issue is that a simple ROA measure does not distinguish between different classes of assets with varying levels of risk. Recall that balance sheet assets are typically book value based and not market value based.

Early attempts to address these issues focused on shifting to performance measures that are defined relative to capital rather than to assets. This approach addresses the first fundamental issue as return on equity (ROE) captures the impact of financial leverage as well as, in theory, the impact of non-balance sheet assets. Both EC models and regulatory capital models attempt to address the second issue by focusing on market valuation directly, as in the case of economic models, or by adjusting book-value measures, as in the case of regulatory models.

The objective of a risk-adjusted performance measure is to define a consistent metric that spans all asset and risk classes, thereby providing an ‘apples to apples’ benchmark for evaluating the performance of alternative business opportunities. RAPMs thus become an ideal tool for capital allocation purposes.

RAPMs come in many different forms, but they can all be loosely defined as a return on capital whereby the measurement of asset riskiness is a key component of the derivation of the formula. In most of the more sophisticated applications of RAPM, asset riskiness is modelled explicitly with respect to a distribution of default-adjusted returns directly, or with respect to a distribution of default losses that is then netted against nominal returns on assets. From these distributions, expected, worst-case, and $x^0\%$ confidence interval default losses (or default-adjusted returns) can be defined and applied to either the return measures or the underlying capital measure.

Two broad classes of RAPM measures include *risk-adjusted return on capital* (RAROC) and *return on risk-adjusted capital* (RORAC). The former applies the risk adjustment to the numerator while the latter applies the risk adjustment to the denominator. Often the distinction between these approaches becomes blurred, with risk adjustments occurring in both the numerator and the denominator, prompting the increasing usage of the term *risk-adjusted return on risk-adjusted capital*

(RARORAC). Nonetheless, common to all these approaches is the principle of incorporating the joint default likelihood of a bank’s obligors explicitly into a bank’s RAPM.

III.0.5.2 Mechanics of RAROC

All RAROC models follow the simplified general formula

$$RAROC = (revenues - costs - expected losses) / capital. \quad (III.0.17)$$

Revenues include all the nominal returns on assets, while costs include all returns to the liabilities holders of the bank. Note that, in practice, all other sources of revenue, including service fees, and all other sources of costs, including general overhead, would normally be incorporated into the RAROC measure. Expected losses would be determined by a risk assessment of the asset base, capturing losses arising from all sources, including credit risk, market risk and operational risk.

The example in the introduction considered a bank whose only activities were the taking in of deposits and the extending of credit. In that case the above equation can be rewritten as,

$$RAROC = (A_0 r_A - D_0 r_D - expected losses) / EC_0. \quad (III.0.18)$$

While a RAPM measure like RAROC is certainly a better indicator of a firm’s overall performance relative to other firms than a more traditional ROA or ROE approach, the true benefit for an individual firm is that it provides a consistent metric to evaluate the performance of the firm’s portfolio of assets, regardless of their unique levels of risk. Taken one step further, it also provides a benchmark for making allocation decisions, while appropriating scarce capital amongst possible new investment opportunities.

Example III.0.3: A Simple Model for RAROC

We return to Example III.0.1, with a firm with liabilities of $D_0 = \$92$ million in deposits at a cost of debt of $r_D = 5\%$, which have been invested in $A_0 = \$100$ million of assets (40% at a nominal return of 6.75% and 60% at a nominal return of 7%). The weighted average nominal return across the \$100 million in total assets is $r_A = 6.9\%$ (a compounded spread of 1.81%). The current capital for this firm was calculated as $E_0 = \$8$ million.

The firm’s balance sheet can be conceptually decomposed into two balance sheets, one associated with each asset class. The amount of debt each asset class can support is determined by the amount of EC that must be allocated to that asset class. Assuming the total worst-case loss in value associated with asset class 1 is 18%, then the economic capital, $EC_{1,0}$, associated with asset class 1, $A_{1,0}$, can be determined in a similar manner to the EC for the entire firm:

$$\begin{aligned}
 EC_{1,0} &= A_{1,0}\{1 - (1 + r_{1,A})(1 - l)/(1 + r_D)\} \\
 &= 60\{1 - 1.07(1 - 0.18)/1.05\} = 9.86.
 \end{aligned}$$

Likewise, assuming the total worst-case loss in value associated with asset class 2 is 10.5%, the EC, $EC_{2,0}$, associated with asset class two, $A_{2,0}$, can be determined from equation (III.0.3) as

$$\begin{aligned}
 EC_{2,0} &= A_{2,0}\{1 - (1 + r_{2,A})(1 - l)/(1 + r_D)\} \\
 &= 40\{1 - 1.0675(1 - 0.105)/1.05\} = 3.60.
 \end{aligned}$$

For illustrative purposes, the current balance sheet can therefore be decomposed on the basis of asset class as shown in Table III.0.2.

Table III.0.2: Current balance sheet

	Asset Class 1	Asset Class 2	Total
Assets	60	40	100
Debt	50.14	36.4	86.54
EC	9.86	3.8	13.46

Table III.0.3 illustrates the expected balance sheets for each asset at year end and, thus, the RAROC or the return on EC for each asset can be measured by the change in the equity position over the year (RAROC in is calculated as $EC_1/EC_0 - 1$).

Table III.0.3: Expected balance sheet

	Asset Class 1	Asset Class 2	Total
Assets	63.2	42.16	105.36
Debt	52.64	38.22	90.86
EC	10.56	3.94	14.5
RAROC	7.01%	9.5%	7.68%

Of course, the change in equity must reconcile with the more familiar relationship,

$$\begin{aligned}
 EC_1 - EC_0 &= (A_1 - D_1) - (A_0 - D_0) \\
 &= A_0 r_A - D_0 r_D - \text{expected losses}.
 \end{aligned}$$

In this example, while asset 1 has the higher nominal return, its RAROC is in fact slightly lower than that of asset 2 as proportionately more EC must be allocated to it to compensate for its higher risk. In other words, the excess nominal return of asset 1 over asset 2 is not quite enough

to compensate for its increased risk and therefore, on a risk-adjusted basis, asset 1 is a less desirable investment.

Note that in this simple example the sum of the EC of each asset class equals the EC of the firm as a whole. This implies that no diversification exists between the two asset classes because the risk of each asset class on a stand-alone basis is equal to each asset class's *contribution* to the overall risk of the firm.

III.0.5.3 RAROC and Capital Allocation Methodologies

When RAROC is used to measure the performance of an asset classes or business, or to allocate capital, it is important to highlight that the denominator measures the *capital contribution* of the asset or business to the overall portfolio. Hence, it is directly linked to the asset allocation methodology chosen by the institution (see previous section). Thus, for example, if a firm uses stand-alone risk contributions, the measure of performance will not account for the diversification opportunities that a given asset or business brings to the overall portfolio. In general, it is beneficial to use allocation methods that account for diversification, such as the marginal risk contributions.

Example III.0.4:

In the simple example above, the capital contribution of each asset was measured as the capital each asset consumes in the scenario that produces the 'extreme' 1% loss. Both asset classes incur a 12% loss in this scenario. This is actually consistent with the marginal risk allocation methodology.¹⁴ Furthermore, in this simple example, the marginal contributions coincide with the stand-alone contributions given the discrete nature of the problem and the high correlation of the asset classes implied by the scenarios.

III.0.6 Summary and Conclusions

The primary role of capital in a firm, apart from the transfer of ownership, is to act as a buffer to absorb large unexpected losses, protect depositors and other claim holders, and give external investors and rating agencies enough confidence in the financial health and viability of the firm.

We distinguish between three different types of capital: the actual, physical capital that a firm holds (book capital); the economic capital associated with a targeted level of solvency and

¹⁴ One can show that for quantile-based measures such as VaR, the derivative in equation (III.0.16), which leads to the marginal capital allocated to a given asset class (or sub-portfolio), is given by the expected losses of that asset conditional on the total portfolio losses being equal to VaR – that is, the expected losses corresponding to all scenarios which lead to the given VaR (see Gouriéroux *et al.*, 2000).

assessed through the use of internal models; and, for banks, the minimum capital imposed by regulatory authorities (regulatory capital). In practice, many firms hold book capital in excess of the required economic and regulatory capital. This reflects some historical and practical business considerations and a more conservative view on the applicability of the models.

Economic capital is a powerful business management tool, since it provides a consistent metric for risk aggregation, performance measurement, and asset and business allocation. The objective of a risk-adjusted performance measure is to define a consistent metric that spans all asset and risk classes, thereby providing an ‘apples to apples’ benchmark for evaluating the performance of alternative business opportunities. Economic capital management tools generally require a bottom-up approach for its estimation, in order to support a practical, risk-sensitive capital allocation methodology.

Regulatory capital and economic capital have differed substantially in the past, particularly for credit risk (and also regulatory capital under Basel I did not cover operational risk). The new Basel II Accord for banking regulation has introduced a closer alignment of regulatory capital with economic capital and current best-practice risk management by introducing operational risk capital and allowing the use of internal models for both credit risk and operational risk. This results in minimum capital requirements that are more risk-sensitive. Finally, with its three-pillar foundation, Basel II focuses not only on the computation of regulatory capital, but also on a holistic approach to managing risk at the enterprise level.

References

Artzner, P, Delbaen, F, Eber, J-M, and Heath, D (1999) Coherent measures of risk, *Mathematical Finance*, 9(3), pp. 203–228.

Basel Committee on Banking Supervision (1988) International convergence of capital measurement and capital standards. Available at <http://www.bis.org>

Basel Committee on Banking Supervision (1995) Basel capital accord: treatment of potential exposure for off-balance-sheet items. Available at <http://www.bis.org>

Basel Committee on Banking Supervision (1996) Overview of the amendment to the capital accord to incorporate market risk. Available at <http://www.bis.org>

Basel Committee on Banking Supervision (2003a) The new Basel capital accord: Consultative document. Available at <http://www.bis.org>

Basel Committee on Banking Supervision (2003b) Trends in risk integration and aggregation. Working paper, available at <http://www.bis.org>

Basel Committee on Banking Supervision (2004) International convergence of capital measurement and capital standards: A revised framework. Available at <http://www.bis.org>

- Denault, M (2001) Coherent allocation of risk capital, *Journal of Risk*, 4(1), pp. 1–34.
- Gouriéroux, C, Laurent, J-P, and Scaillet, O (2000) Sensitivity analysis of values at risk', *Journal of Empirical Finance*, 7(3–4), pp. 225–245.
- Kalkbrenner, M, Lotter, H, and Overbeck, L (2004) Sensible and efficient capital allocation for credit portfolios, *Risk*, pp. S19–S24.
- Koyluoglu, H, and Stoker, J (2002) Honour your contribution, *Risk*, April, pp. 90–94.
- Kupiec, P, (2002) Calibrating your intuition: Capital allocation for market and credit risk. Working paper 99/02, IMF, available at <http://www.gloriamundi.org/picsresources/pkcyi.pdf>
- Matten, C (2000) *Managing Bank Capital*. Chichester: Wiley.
- Mausser, H, and Rosen, D (2004) Scenario-based risk management tools. In S W Wallace and W T Ziemba (eds), *Applications of Stochastic Programming*. Philadelphia: SIAM.
- Merton, R C, and Perold, A F (1993) Theory of risk capital in financial firms, *Journal of Applied Corporate Finance*, 6(Fall), pp. 16–32.
- Perold, A F (2001) Capital allocation in financial firms. Harvard Business School Working Paper 98-072, available at http://papers.ssrn.com/paper.taf?abstract_id=267282
- Praschnik, J, Hayt, G, and Principato, A (2001) Calculating the contribution, *Risk*, 14(10), pp. S25–S27.
- Saita, F (2003), Measuring risk-adjusted performances for credit risk. Working Paper 89/03, March, available at <http://www.sdabocconi.it/it/ricerca/pubblicazioni/dir2003.html>
- Smithson, C (2003) Economic capital - how much do you really need?, *Risk*, November, pp. 60–63.
- Tasche, D (2000) Conditional expectation as quantile derivative. Working paper, Technische Universität München.
- Tasche, D (2002) Expected shortfall and beyond. Working paper, Technische Universität München.

The Professional Risk Managers' International Association



PRM Program Self-Study Guide

Updated: March 17, 2003

Produced by Value Consultants Limited
© 2003, Value Consultants Ltd. / Professional Risk Managers' International Association

Table of Contents

<i>Table of Contents</i>	2
<i>Introduction</i>	3
<i>Study time frame</i>	6
<i>Reading list</i>	6
Exam I.....	7
Exam II	8
Calculus	8
Linear Algebra	8
Probability.....	8
Exam III	9
Market Risk	9
Credit Risk	9
Operational Risk	10
Exam IV	10
Barings.....	10
Metallgesellschaft	10
Long Term Capital Management	10
Group of 30 Report	11
PRMIA’s standards.....	11
<i>Exam I</i>	12
Finance theory.....	12
Portfolio theory and asset pricing	12
Contingent claims	15
Financial instruments.....	18
Markets	29
<i>Exam II</i>	32
Calculus	33
Linear algebra	36
Probability and statistics	40
<i>Exam III</i>	47
Market Risk.....	47
Credit Risk	56
Operational Risk	66
<i>Exam IV</i>	71
Case Studies.....	71
PRMIA Bylaws and Code of Conduct.....	75
<i>Appendix: further reading</i>	77

Introduction

Congratulations for deciding to sit the Professional Risk Manager examination. This study guide is designed to help you prepare efficiently for the program.

This document complements the study material and syllabus available from PRMIA (also obtainable on www.prmia.org, menu: Certification, Candidate Information, Exam Guide). Here we point the user to useful resources and indicate what effort should be put in each part of the syllabus. It is aimed at assisting candidates in the preparation for the exam.

The Risk Management profession has changed a lot in the last few years. We expect it to change a lot, too, in the next few years and beyond. This is why the present exam has been created: to ensure that the Risk Managers of tomorrow have the competences and ethics to equip the financial institution with the right tools to manage risks. The Risk Management profession, since its start, has always been fascinating and remunerative. It will no doubt continue to be so, and to offer tremendous prospects for research and growth. We hope, if you are not already a Risk Manager, to welcome you in to this profession. We wish you every success with the exam and, perhaps more importantly, hope that you learn useful skills and concepts from your studies that you can apply to your work...

More and more training courses are being offered round the world. Some candidates have also organised self study groups, in which candidates meet after work, for each participant to present on their field of expertise and to solve together sample exam questions. The total time needed to study varies according to previous experience, but few people, if any, regardless of their expertise, claimed to have passed the exam without some specific self-study. The exam program represents a serious commitment in time and effort.

Whether the candidate chooses to study alone exclusively or within a study group, the present study guide can help to orient the effort and to facilitate research.

The reading list given below is to be regarded as the minimum required. We shall refer to this short reading list extensively throughout this guide. There is a further more comprehensive reading list given at the end of this document. These lists are by no means exhaustive, and should be completed by other sources and personal on-the-job experience. Candidates may succeed if they do not read all, but at their own risks!

The PRM exam program is split into four modules. These can be taken separately (i.e. at different times) and in any order, or may be taken as one complete exam. This provides you with a great deal of flexibility in preparation. You may wish to focus on the exam or exams that will be most challenging to you and take those separately and last. Some prefer to get the toughest one out of the way first. We recommend, for most, though, that you study and take the examinations in the order given below:

EXAM MODULE	TOPICS	DURATION	PERCENTAGE OF TOTAL POINTS
1	Finance Theory, Financial Instruments and Markets	90 minutes	25%
2	Mathematical Foundations of Risk Measurement and Pricing	2 hours	20%
3	Risk Management Practices	90 minutes	30%
4	Case Studies PRMIA Standards of Best Practice Conduct and Ethics, PRMIA Bylaws	1 hour	25%
Full PRM Exam	All Subjects	6 hours including a one-hour break	100%

The modules are not of the same length, and the weightings of the modules are not equal. As an example, the most important module is module 3, which carries 30% of the questions for the PRM qualification, yet the exam lasts 30 minutes less than module 2, which carries only 20% of the questions. Bear in mind that it is vitally important to finish each module completely in the time allotted. Do not linger over questions longer than is sensible. For example, if the exam has 30 questions in 90 minutes, do not spend longer than three minutes per question. If at the end of three minutes you have not answered the question, guess the best answer you can (ignoring the obviously wrong), mark your answer and move on. If you do have any spare time at the end of the exam (unlikely!) you can always go back and review the answer. However, make absolutely sure that you have an answer for every question at the end of the exam! As the exam is delivered via computer, marking and review of questions is very simple.

The mathematics needed for the exam generally goes beyond what is needed in finance day-to-day, but are similar in level to that required in basic university-level courses. This is to insure that successful candidates are equipped for the quantitative challenges of tomorrow.

It is often advisable to try to solve the questions by using intelligent shortcuts instead of getting into the full algebra of the solutions. For example, if the question is, how much is square root of 23409?, with solutions being: -123, 152, 153 and 155, no use to rush on a calculator, as the last digit indicates that 123 and 153 are the only possible solution (as only 3*3 will give us a 9 at the end). -123 can be eliminated because it 'looks' too small ($12(0)*12(0)=144(00)$). Similarly in trading situations, one must be able to have an intuitive idea of the solutions and cut through the analytics by rough mental calculations. If a question takes too many messy calculations, this is a sign that you had better save your brains and find a smarter way to solve it. Frequently some of the answers can be

dismissed fairly easily since they can be ruled out using simple logic, often leaving two similar answers that need a bit more work to identify the correct solution.

The next sections are organised as follows: preliminaries, expanded topics from the syllabus, self-study resources, and Study Questions with indications of solutions.

The reading times are only indicative, and can vary very widely according to previous experience on the topic matter. As a rough guideline, a skim-read should take no more than one minute per page; a full reading should take two minutes per page, and for heavily mathematics-oriented topics three minutes per page.

Study time frame

The table below gives an estimate of the required study time (in minutes) for each module and topic. Note that the time spent on the mathematics section can be greatly reduced if the candidate is already familiar with these topics.

	Minutes	Totals	% total
Exam/Module I		2659	32%
Finance Theory	195		
Contingent Claims	145		
Financial Instruments	1154		
Markets	1165		
Exam/Module II		3066	37%
Calculus	965		
Linear Algebra	666		
Probability	1435		
Exam/Module III		2105	26%
Market Risks	1225		
Credit Risks	655		
Operational Risks	225		
Exam/Module IV		380	5%
Case Studies	290		
PRMIA Standards	90		

The total time required is therefore approximately 140 hours.

Reading list

There are ten texts listed here. They form the basic reading list for the course. The books should be readily available from most business bookshops or online vendors. You can click on each image to order through Amazon.com. If you need to buy all of these, the total cost should be approximately USD 425 or GBP 275. Consider sharing within a group or looking for used texts to reduce costs. After the first full reference to the text, we shall refer to the book by a shortened reference given below.

The chapter references given below contain all the topics within the syllabus; specific references to sub-topics are given in the detailed section of this guide. There are two ways of using this guide: 1) By reading the texts chapter by chapter as given below, referring to specific topics as the relevant chapter is read, or alternatively 2) By following the detailed syllabus and reading the relevant topic as they arise. The later method should be faster and more efficient, but you may lose some of the context and continuity.

Some of the books referenced are of Schaum's Outline series. For those unfamiliar with these texts they offer a very condensed summary of topics with very extensive worked examples, reading/following through the worked examples is a very important part of the preparation process and should be regarded as an integral part of the chapter.

Exam I

Finance Theory, Financial Instruments and Markets



Bodie, Zvi, Alex Kane and Alan J. Marcus, *Investments*, 5th Edition, New York: McGraw Hill

- Finance Theory Chapters 6, 7, 8, 9, 10, 11 and 12
- Financial Instruments Chapters 14, 15, 21 and 23
- Markets Chapters 1, 2, 3, 20 and 22

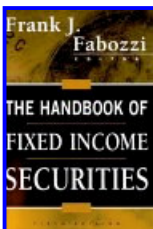
Referenced as: Bodie



Hull, John C., Prentice Hall, 5th Edition, January 2002

- Finance Theory Chapters 7, 8, 10, 11, 12
- Financial Instruments Chapters 3, 6, 7, 9, 13, 25, 27 and 29
- Markets Chapters 1, 2, 5 and 7

Referenced as: Hull



Fabozzi, Frank, *Handbook of Fixed Income Securities*, 6th Edition, McGraw-Hill, 2000

Financial Instruments Chapters 1-4, 8, 9, 11, 14 and 24

Referenced as: Fabozzi



Reuters: *An Introduction to Commodity, Energy and Transport Markets*, John Wiley & Sons Finance

Referenced as: Commodities

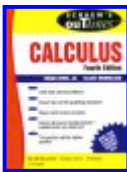


Reuters: An Introduction to Foreign Exchange and Money Markets, John Wiley & Sons Finance
Referenced as: FX

Exam II

Mathematical Foundations of Risk Measurement and Pricing

Calculus



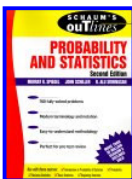
Schaum's Outline of Theory and Problems of Calculus, Ayres and Mendelson, 4th Edition, 1999 Calculus Chapters 1-14, 17-18, 23, 29-36, 46-49, 52, 54 and 56. The early references here (Chapters 1-14) cover fairly elementary topics and act as general mathematical preparation.
Referenced as: Calculus

Linear Algebra



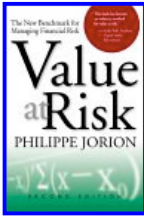
Schaum's Outline of Linear Algebra, Lipschutz and Lipson, 3rd Edition, December 6, 2000
Linear Algebra Chapters 1-4, 8 and 9
Referenced as: Linear Algebra

Probability



Schaum's Outline of Probability and Statistics, Spiegel, Scholler, Srinivasan and Srinivasan, 2nd Edition, March 17, 2000

Chapters 3, 5, 7, 8 and 10
Referenced as: Probability



Jorion, Philippe, Value at Risk: The New Benchmark for Managing Financial Risk, 2nd Edition, McGraw Hill, 2nd Edition
Probability Chapters 4 and 12
Referenced as: Jorion

Exam III

Risk Management Practices

Market Risk

Probability
Chapters 4

Fabozzi
Chapter 5

Jorion
Chapters 3 (Section 3.3), 5-11, 14 and 21 (Section 21.1)

Hull
Chapters 5 (duration and convexity analysis section), 14-17



Crouhy, Galai and Mark, Risk Management, McGraw Hill 2001
Chapters 1, 5 and 6 including appendices
Referenced as: Crouhy

Linsmeier and Pearson. Risk Measurement: An Introduction to Value at Risk.
www.gloriamundi.org/picsresources/LandP.pdf
RiskMetrics Technical Document Chapters 4, 5 and 6.
<http://www.riskmetrics.com/rmcovv.html>
Return to RiskMetrics - The Evolution of a Standard. Chapters 1, 2 and 3.
<http://www.riskmetrics.com/r2rovv.html>

Credit Risk

Jorion Chapters 3 (Section 3.4) and 13

Crouhy Credit Risk Chapter 7, Chapter 9 Section 5 pp. 368-389 "The KMV Approach" and Chapter 12

CreditMetrics Technical Document, pp. 1-40 and 57-93.
<http://www.riskmetrics.com/cmtdovv.html>

Bank of International Settlements. Credit Risk Modelling Current Practices and Applications. Basle Committee on Banking Supervision, April 1999. pp. 1-33.
www.bis.org/publ/bcbs49.pdf

Bank of International Settlements. Principles for the Management of Credit Risk - Consultative Paper. Basle Committee on Banking Supervision. July 1999. pp. 3-4 and Appendix. www.bis.org/publ/bcbs54.pdf

Bank of International Settlements. "Standardized Approach to Credit Risk. January 2001. pp. 1-52. www.bis.org/publ/bcbsca04.pdf

Operational Risk

Jorion Chapters 18-20
Crouhy Chapter 13

Basle Committee on Banking Supervision. Consultative Document Operational Risk. pp. 1-14. www.bis.org/publ/bcbsca07.pdf

Exam IV

Case Studies PRMIA Standards of Best Practice Conduct and Ethics, PRMIA Bylaws

Barings

Jorion Chapters 2 (Section 2.2.1) and 21 (Section 21.1.2)

Bank of England. Report on the Collapse of Barings Bank, 1995.
www.numa.com/ref/barings/bar00.htm

International Financial Risk Institute. Not Just One Man-Barings.
<http://newrisk.ifci.ch/137550.htm>

Metallgesellschaft

Jorion Chapter 2 (Section 2.2.2)

Digenan, Felson, Kelly and Wienert. Metallgesellschaft AG: A Case Study.
www.stuart.iit.edu/fmtreview/fmtrev3.htm

Krapels, Re-examining the Metallgesellschaft Affair and its Implication for Oil Traders. http://www.esai.com/pdf/Re-Examining_the_Metallgesellschaft_Affair.pdf

Long Term Capital Management

Jorion Chapters 14 (Section 14.4) and Chapter 21 (Sections 21.1.2 and 21.4)



United States Treasury. Hedge Funds, Leverage and Lessons from Long Term Capital Management, pp. 1-42.

www.treas.gov/press/releases/reports/hedgfund.pdf

Shirreff. Lessons From The Collapse Of Hedge Fund, Long-Term Capital Management. <http://newrisk.ifci.ch/146480.htm>

Group of 30 Report

Jorion Chapter 2 (Section 2.3.1) and 21 (section 21.1.1)

Group of Thirty, Derivatives Practices and Principles. www.group30.org/app.htm for purchase, or summary at <http://newrisk.ifci.ch/136160.htm>

PRMIA's standards

Professional Risk Managers' International Association. Standards of Best Practice, Conduct and Ethics (2002). www.prmia.org/pdf/Conduct.PDF

Professional Risk Managers' International Association. Bylaws (2002). www.prmia.org/pdf/Bylaws.PDF

Exam I

This exam contains: Finance Theory, Financial Instruments and Markets. It lasts 90 minutes, and counts for 25% of the total exam points. The total time estimated to read the source texts for this module is 44 hours.

Finance theory

This part includes 40% of the points of exam module I, and therefore 10% of the marks for the entire PRM qualification.

Portfolio theory and asset pricing

Mean-Variance analysis

What this is about

An asset, and a portfolio, can be characterised by its expected return and its risk, measured by its standard deviation or variance (variance is standard deviation squared). These are interrelated, as riskier assets tend to offer higher expected returns. MV analysis uses the historical relationships of risk / return for market instruments to forecast future risk / return and to evaluate new instruments relative to existing products.

Readings

Bodie

Chapters 6 (20 min), 7 (15 min), 8 (30 min)

Key things you should know

Risk aversion, risk preferences

Risk-free asset

Stock price return and volatility

Expected returns, standard deviation and variance

Portfolios of assets

The capital market line

Correlation of returns

Study Questions

Q. When does diversifying an investment from a single asset into a portfolio show more merit?

Considering two assets, provided they are not perfectly correlated, the loss on the underperforming asset will tend to be compensated for by the gain on the other. This will therefore lower the total risk (i.e. the P&L variance of the basket of assets). In more quantitative terms, the variance of the return of a two-asset portfolio is lower than the sum of the variances provided that they are not perfectly correlated. Hence the total portfolio will be less risky than individual assets. This is more valid when the asset returns are less correlated, and even more when the assets are negatively correlated. This is why investing in gold-related assets, when gold was a hoarding tool, was a good defensive measure: when the political situation went bad, most assets depreciated while scared investors took their savings into gold, waiting for a war.

Capital Asset Pricing Model

This part includes APT, ICAPM and CCAPM.

What this is about

Finance theory has been hugely developed in the last few decades, discovering arbitrage-based models that help to quantify the relative prices of assets. The basic Capital Asset Pricing Model (CAPM) has been complemented by different models relaxing the underlying assumptions, and using different representations of time. Basically, CAPM relates the market rate of return to risk. The candidate needs not understand the different versions of the framework and how they can best apply.

Readings

Bodie chapter 9 (40 min), 10 (25 min)

Key things you should know

Arbitrage-free pricing
The efficient frontier
The market portfolio
Determining the optimal portfolio
Correlation between assets
Capital Asset Pricing Model
CAPM with borrowing costs

Study Questions

Q. What would make a knowledgeable investor choose a CAPM-suboptimal portfolio?

Apart from idiosyncratic preferences (ethical investing or others), a knowledgeable investor will have views on the market that do not necessarily coincide with the views of the market. He will then see different expected returns, assess different variances. Moreover, he will have some practical difficulties to invest into the full market portfolio.

Q. Assume you live in a CAPM world and the expected return on the market portfolio is 9%, while the risk-free rate is 3%. If the beta of stock A is 1.3, the expected return on A is:

- a) 14.7%
- b) 12.9%
- c) 10.8%
- d) 16.8%

CAPM says that the expected return is the risk-free rate plus beta times the market premium (which is the difference between the market expected return and the risk-free rate, here 6%): $3 + 1.3 * 6 = 10.8$: c)

Efficient frontiers, capital market line, beta

What this is about

Assuming variances, correlations and expected returns of assets are known, one can construct an optimal portfolio for every investor, taking into account his risk preferences, using efficient frontier analysis and betas. Through the capital market line, one can associate a given degree of acceptable risk to a given return.

Readings

Bodie, chapter 11 (15 min) and 12 (20 min)

Key things you should know

Determinants of the capital market line
Market price of risk
Stock betas
The Index model
Levels of market efficiency
Market anomalies

Study Questions

Q. What would happen if an investor's preference curve were not tangent to the capital market line?

An investor's preference curve is the (convex) curve representing the points of same degree of satisfaction with the risk and return of a portfolio. If the investor is at a point on his curve which is below the capital market line, he will move his preferences until he touches the line, and will then move along the line to another, more ambitious preference curve. In other words, he would raise his expectations. If no point of his preference curve touches the capital market line, i.e. he has too greedy expectations, he would get into lower preference curves, parallel to the first one, until a point of a preference curve touches the market line. In short, this investor will correct his expectations to fit the market line.

Tobin approach

What this is about

This argues that the main thing that matters about a company's value, over the long term, is at what cost the company can be replaced? (Tobin's q coefficient): this is commonly known as book, or intrinsic value.

Readings

Bodie, chapter 18 (30 min)

Key things you should know

Balance sheet structure
Book Value
Main valuation methods of company assets
Liquidity, quick ratio, acid test

P/E ratio
Equity valuation

Study Questions

Q. How can Tobin's approach explain the dot-coms debacle?

Quite well. The main assets of most dot-com companies are (were?) the present value of growth opportunities (PVGO) based upon the Internet, rather than their few tangible assets. PVGO, as a volatile asset, needs to be cared for by vision and business plans with a very good future visibility. When growth opportunities do not show up, the market is merciless in its reassessment of PVGO.

Contingent claims

The Black-Scholes-Merton model

What this is about

The Black-Scholes-Merton model allows the pricing of an option in relationship to its underlying instrument, using its volatility, owing to an arbitrage relationship, in continuous time. Assuming that an instantaneously hedged position of the derivative and its underlying asset can be constructed and rebalanced continuously, the model allows the evaluation of the total hedging costs of the portfolio. This risk-free portfolio must earn the risk-free rate in order to avoid riskless arbitrage, hence the costs of hedging the portfolio can be established. The model then relates the cost of replicating the risk of the derivative to the value of the derivative itself, which by arbitrage must be equal. The model therefore prices the derivative by modelling the cost of exactly hedging the derivative.

Readings

Hull Chapter 12 (30 min), 14 (20 min)

Bodie chapter 21 (40 min)

Key things you should know

Criteria for accepting/rejecting a model

Assumptions of the model

Random walk hypothesis

Delta-neutral position

Dynamic hedging

The 'Greeks' and their behaviour

Black-Scholes-Merton formula

Study Questions

Q. How critical is it to apply the Black-Scholes framework to European-type options only?

In theory, it is not at all, as exercising an American type or Bermudan-type option before maturity is renouncing the volatility value of the option (subject to the presence of dividends). A holder of such option should then only hedge his position, or sell the option if it is a negotiable one. In practice, the difference matters, as a non-European option offers more flexibility to its owner. This flexibility can be very valuable in the presence of difficulties to hedge, uncertain

dividends, or frictions in the market. The effects of early exercise can substantially increase the value of derivatives over their European equivalent in certain circumstances (e.g. deep in the money put on a stock paying a large dividend near expiry). The basic approach to valuation however does not change from the Black-Scholes framework, but it is important to solve the partial differential equation (PDE) in a numerical framework which allows the effects of early exercise to be modelled, e.g. in the Binomial model rather than using the regular BS "closed form solution".

The Binomial model

What this is about

Assuming the underlying asset price of an option as going up or down a certain percentage each finite time period, a pricing tree can be built, allowing an option pricing model in discontinuous time. Essentially, this model is a numerical grid for the solution of the Black-Scholes PDE and uses a stock price tree extending from today to the expiry date of the option. The value of the option is obtained at the maturity boundary (i.e. just the local intrinsic value) and this is used successively to step back to an early node in the binomial tree, taking into account the risk-free discount rate and probabilities (risk-neutral) of price movement in the underlying stock price. By working back step-by-step, the model "solves" the BS PDE and gives us the "fair value" for the option. By analysing the value of the option exercised or not at each point in the tree, it is possible to include "path-dependent" (such as American exercise) features in the pricing model.

Readings

Hull Chapter 10 (20 min)

Key things you should know

Risk-neutral valuation

Construction of a binomial tree, probabilities, step size, u , d , and p

Discrete time models

How to incorporate path-dependent features

The effects of dividends

Recombining price trees

Study Questions

Q. Which of the following is true:

- a) Non-Markovian interest rate processes are usually represented by recombining trees
- b) Markovian interest rate processes are usually represented by recombining trees
- c) Non-Markovian interest rate processes are usually represented by trinomial trees
- d) none of the above

Markovian processes are stochastic processes where successive prices are partially determined by previous prices with the size of price movement also proportional. They can be, and usually are, represented by a recombining tree: b)

Put-call parity

What this is about

Based upon arbitrages, we can find out the price relationship between a call and a put on the same asset at the same strike price and maturity, buying the call, shorting the put, short-selling the share and borrowing or lending the cash. Synthetically, we have created the risk profile of the underlying asset out of a Call and a Put and therefore the value of this synthetic asset must be equal to the underlying asset, else risk-less arbitrage can occur.

Readings

Hull, chapter 8 section 4 (10 min)

Key things you should know

Put-Call parity relationship
Effect if the options are American exercise
Put on a dividend-paying stock (i.e. parity does not hold)
Put pricing
Option exercise strategies

Study Questions

Q. What should an investor do to hedge a position including a long call and a short put at the same strike price?

A graphic representation shows that this is equivalent to a long position in the underlying. To hedge this, just short (sell) the underlying and invest the cash at the risk-free rate until the expiry date of the options.

Interest rate parity

What this is about

Given a forward price for a FX transaction and the interest rate for one currency to the forward date, arbitrage arguments allow the evaluation of the interest rate in the other currency. Hence the relationship between forward FX prices and the relevant interest rates is called the interest rate parity.

Readings

Bodie, chapter 23 section 1 (15 min)

Key things you should know

FX spot and forward
Interest rates
Parity theory

Study Questions

Q. What should an operator do if $JPY/USD = 120$, forward 6M = 114, 6M interest for USD and JPY are 4% and 1% respectively?

The forward price suggested by the market is: $120 \times 1.005 / 1.02 = 118$. The interest rate in JPY is undervalued. Hence borrow JPY at 1%, convert these into USD spot (120) and lend at 4%, then hedge your spot position by selling forward USD for JPY at 1USD = 114 JPY. These trades will lock-in a risk-free arbitrage profit of JPY 4 at expiry.

Cash-and-carry pricing

What this is about

Given the possibility to buy spot and store an asset, a relationship can be established between spot and forward prices given the level of interest rates and storage costs / dividends. Effectively, this is the arbitrage forward price.

Readings

Hull, pp. 2-5 (10 min)

Key things you should know

Cost of carry/storage
Forward prices

Study Questions

Q. The 1-year forward price of a commodity is 3% above the spot price. What should the cost of storage be to deter arbitrage, knowing that one-year interest is 2%? No margins are to be considered.

To have an arbitrage-free price in the physical commodity in 1 year, we should be economically indifferent to buying it spot and storing it or buying it forward. Hence the cost of financing (interest) plus the cost of storage should equate the forward premium, hence the cost of storage should be of $3 - 2 = 1\%$.

Financial instruments

A descriptive knowledge and knowledge of pricing mechanisms will be tested. This part accounts for 20% of the points of exam module I and hence 6% of the entire PRM exam marks

Compounding methods

These include simple, annual, semi-annual and continuous methods.

What this is about

The equivalent interest rates according to the frequency of payment of interest can be derived by a simple formula. The amount of interest payable for a given rate is dependent on the date count convention and compounding frequency used hence 10% semi-annual on a 30/360 day count is a different amount of interest payable than 10% annually on an actual/actual day count convention. Different markets such as bonds and loans have different conventions. Continuous compounding is most often used in financial modelling as it makes manipulation of equations much easier.

Readings

Hull Chapter 5 (pp. 93 – 119) (60 min)

Key things you should know

How to calculate interest payments depending on compounding frequency and day count convention

How to convert from one convention to another

Standard money market and bond market conventions

Study Questions

Q. A bond pays interest of 10% semi-annually on an actual/actual convention, what is the rate of interest on an annual actual/actual basis?

The day count convention is the same so we do not have to worry about converting this, payment of 10% semi-annually means payment of 5% each half year, so this equates to $(1+5\%)*(1+5\%)=(1+10.25\%)$ i.e. 10.25% on an annual compounding basis.

Bonds

Knowledge of simple (non-optional) bonds will be tested.

What this is about

The most classical fixed-income instruments are straight bonds, which, against payment of a principal, provide a regular coupon (interest) and a redemption payment. Bonds are issued by companies to raise money to finance their business and are issued via Banks to investors. Bonds are a type of tradable loan that are characterised by a face or notional amount (usually USD10,000) for each bond and a coupon (interest rate) payable. Bonds can have fixed or floating interest rates, the later known as FRNs (floating rate notes). The prices of bonds are usually quotes as a percentage of face (or notional) value e.g. 101%.

Readings

Fabozzi, Chapter 4 (pp. 51 – 83) (40 min)

Bodie, chapter 4 (60 min)

Key things you should know

What the terms face value, notional, coupon, redemption, maturity and principal mean

How bonds are used raise capital

How bonds are traded between investors via dealers

The inverse relationship between bond price and yield

Simple concepts of yield to maturity

Credit spreads

Study Questions

Q. A fixed-income instrument will pay 12-month Libor on a 1,000 Swiss Francs (CHF) face value two times: one year from today and two years from today (no principal payment). The rates are set in arrears (payments at the end of a year reflect the Libor rate at the beginning of the year). What is the price of this instrument if the (zero-coupon) two-year CHF swap curve is a 3% for all maturities?

- a) CHF 57.4
- b) CHF 1000
- c) CHF 106
- d) CHF 67.6

Swap rates are quoted against LIBOR (floating rate index) and a flat swap curve implies that forward LIBOR is constant over the curve and equal to the swap rate. The cash flows are, for the first year, 3% on 1000 CHF payable in two instalments, and the same for the second year: 60 CHF in total. As there is no principal repaid, b) and c) are out of the question. As these payments are discounted (as coming later than today), 57.4 is the only possible solution: a)

Floating rate notes

What this is about

Floating Rate Notes (FRNs) are fixed-income instruments (bonds) providing a coupon fluctuating with a given interest rate index (example: Libor) and a redemption payment. FRN typically trade close to par (100) price since the interest rate they pay is usually equal to LIBOR plus a spread appropriate to the issuers' credit risk.

Readings

Bodie Chapter 14 (p. 419) (4 min)

Key things you should know

How the coupon payment is calculated. How the floating index reference (e.g. LIBOR rate) is determined. How price of the FRN is effected by changes in interest rate and credit worthiness of the issuer.

Study Questions

Q. A company issues an FRN at par, which pays LIBOR plus 125 b.p., quarterly. Overtime the credit rating of the firm declines from AA to BBB, how would this effect the price of this FRN?

The spread to LIBOR represents the extra return for assuming the credit risk of the issuer of the bonds. If the creditworthiness of the issuer declines the market will require a higher return for assuming this risk, i.e. they require a higher spread that 125 b.p., so the will pay less for this FRN and it will trade at a price below par (100).

Futures, Forwards

What this is about

Simple buy/sell transactions concluded now for a delivery date in the future can be termed forwards. When these transactions can be done through an exchange and freely tradable, these are futures. An example of a forward contract is an agreement today that one counterparty will deliver an asset (e.g. a barrel of oil) in one year's time for the payment at that time of an amount of money agreed and fixed today. Futures contracts are similar to forwards, but usually involve the payment of daily margins to the exchange based on the mark to market value of the contract and are standardised.

Readings

Hull Chapter 2 (pp. 36 – 37) Chapter 3 (pp. 49 - 52) (20 min)
Bodie chapter 23 (50 min)
Commodities pp. 133-255 (240 min)

Key things you should know

Relationship between spot and forward prices
Cost of carry (arbitrage) calculation
How futures are margined
Contango and positive carry markets

Study Questions

Q. If a dealer sells a security to a retail customer and the customer pledges to resell that security to the dealer, the trade is called:

- a) a reverse repo
- b) a forward sale
- c) a repo
- d) a reverse forward

A 'repo' is a repurchase agreement. It consists of an agreement to sell a security and simultaneously agreeing to buy the same security back at a fixed price at a fixed date in the future. A reverse repo is the agreement to buy and then sell back. As the name indicates, this consists of two operations in one. A forward is an operation concluded now for completion in the future. The character of 'reverse', for a repo, comes from who the active counterparty is and who the 'client' is. A repo is, for an investor, a substitute to a secured deposit. A reverse repo is akin to borrowing a security for the dealer from a client. The deal is therefore a repo for the dealer and a reverse repo for the customer. The question refers to the dealer's perspective. Hence c).

Q. A gas market maker (MM) has agreed to deliver gas at USD 3/MMBtu 6 months from now. The spot price for gas is USD 2.50/MMBtu, the 6-month forward price is USD 2.75/MMBtu, the interest rate is 6% and storage cost is USD 0.03/month per MMBtu. The MM is confident that the price would be USD 2.70 in six months from now. Given the MM's market view, what is the best strategy for the MM to meet its obligation?

- a) Buy on the spot market
- b) Buy the forward
- c) Do nothing

d) MM is indifferent between c) and b)

The reaction of many a candidate will be to consider hedging as a given, and to compare the price of buying spot and storing the gas to the forward price to determine if a), b) or d) is the correct answer. However, the Market Maker has a strong bearish view on the market. In such a situation, there is no point in hedging, but he should just wait for the price to come down to purchase the gas. The fact that hedges are available does not mean that operators should surrender their market views. Hence c).

Options

What this is about

An option is an agreement to provide the right to buy or sell an asset at an agreed price, at or before a certain date in the future. This instrument provides the opportunity for markets to trade the volatility of assets. The option contract is an obligation to the writer and the holder, who pays the premium to buy the option, has the right but not the obligation to use it. Call options are the right to buy and Put options the right to sell. Many kinds of exotic options exist such as barriers or currency protected (quanto) which have additional useful features. Option contracts can be written on virtually any kind of underlying asset.

Readings

Hull Chapter 7, 8 and 9 (pp. 151 – 199) (150 min)

Key things you should know

Basic definitions of Call, Put, strike price, expiry, exercise, volatility and rates
Pay-off diagrams for options and combinations of options such as Call spreads
The Greeks as rate of change of value of option with changes in underlying parameters
How basic Greeks (Delta, Gamma) change depending on the option being in, out or at the money (spot price = strike price)
Exercise types such as European, American, and Bermudan.

Study Questions

Q. In the valuation of derivatives, the expression "change of measure" means:

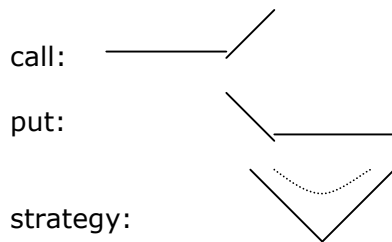
- a. setting the drift to zero
- b. change of volatility
- c. both (a) and (b)
- d. none of the above

The change of measure referred to here is when a real world drift (arbitrage forward growth of the asset price e.g. interest rates minus dividends for a stock) is modified by transforming the drift into other units known as the numeraire (e.g. the numeraire might be the price of a zero-coupon bond rather than say euros) – if the correct numeraire is selected, the risk-adjusted drift of this asset will be zero (i.e. the arbitrage forward price of the asset does not grow with time). However, it does not necessarily mean that the real drift will be zero. Hence the answer is d)

Q. In a long option straddle strategy, where one buys a put and a call simultaneously at the same strike, the following is true:

- a) Delta will be zero, regardless of the level of the spot price
- b) Gamma will be the highest at the money and approaching maturity
- c) Delta will be near to 1 at the money and approaching maturity
- d) Gamma will be zero at the money and approaching maturity

It helps to picture out the payout functions.



The long option strategy will result in the dotting curve above, tending towards the V-shape seen above. Delta (the slope) will not be zero, it could tend to -1 or to 1 . Gamma (the curviness) will tend to infinite when approaching maturity and if at-the-money, when the curve gets into the bottom of the V-shape: b).

Term structure

Knowledge of the basics of term structure modelling will be tested.

What this is about

The relationship between interest rates (or other forward prices such as oil price) at different maturities is called the term structure. This can be modelled so as to determine the right relationship between interest rates at different maturities. One-factor models generally use a short-term interest rate and a stochastic process to model the entire yield curve. This model of yield curve behaviour can then be used to evaluate derivatives based on interest rates. The Black-Derman-Toy option model uses a single factor interest rate term structure model. The term structure can also be modelled using more factors; a two-factor model uses two rate processes, one for short-term rates and the other for long-term rates. The Heath-Jarrow-Morton model is an example of a derivative pricing model that uses a two-factor term structure model. Generally the interest rate term structure has a the feature of mean-reversion so that high rates have a higher probability of declining rather than increasing and low rates have a higher probability of increasing rather than decreasing. The term structure can be represented as $r(t)$ where the relevant interest rate is a function of time.

Readings

Hull Chapter 23 (pp. 537 – 569) (100 min)
 Bodie chapter 15 (30 min)

Key things you should know

- Basic yield curve shapes
- How short-term and long-term rates are related
- Spot and forward rates
- How rate structures can be modelled
- Fitting model parameters to fit current yield curve shapes
- Other term structures such as oil forward curves
- Convenience yield

Study Questions

Q. Assume the following price curve for crude oil in bbl and ignore the time value of money.

Month	Price (USD)
1	21
2	22
3	23
4	24
5	25
6	26

A customer wants a tailored six-month swap with constant volumes, but requests the fixed price for the last two months to be set at USD 20/bbl. What must be the fixed price for the first four months?

- a) not determinable
- b) 26.0
- c) 23.7
- d) none of the above

The question says we can ignore the time value of money, so we do not have to do any present valuing of the cash flows. In this commodity swap the basic idea is to quote one fixed rate (price) per month to supply the commodity on a month basis over the six-month period, hence what we are looking for is a simple volume weighted average. The price of USD 20 for the last two months is below market; hence the other prices must be above market. This does not eliminate any alternative (although b) can look intuitively too high). The sum of prices must equate, hence $(21+22+23+24+25+26)$, which equals the average times the number of prices, hence 23.5 times 6, equals 141. Choose if doing the sum on a calculator is quicker or not. Then $141-20-20$, divided four months, gives 25.25. None of the above: d)

Hybrid instruments

What this is about

Between equity and fixed-income instruments, there exist hybrids, in which rights and remuneration are a combination of both types of instruments. This allows corporations to have a more flexible structure of long-term capital, and creates instruments with different risk profiles that can be tailored to meet the requirements of the issuer and investor. Hybrids often consist of fixed income instruments such as bonds that have an interest rate (or coupon) and/or the redemption amount link to the price of another asset such as a stock index or oil price. So, for example, an oil-linked bond is an example of a hybrid of fixed income and oil risk. Structures with embedded options linked to other risks e.g. default risk of a basket of bonds are also hybrids. Corporates are typically funded by tiers of capital (of which the riskiest is the equity layer), moving up through preference shares to junior debt through to senior debt. Hybrid securities can be created with a risk profile that mixes the risks of the traditional capital structure.

Readings

Bodie Chapter 1 (pp. 1 – 25) (30 min)

Key things you should know

How Hybrids can be broken down into their constituent parts, examples of coupon linked and redemption linked structures, pricing structures by summing the value of component parts, the traditional capital structures of corporations, securities firms and banks.

Study Questions

Q. Which of the following is/are true concerning preferred stocks?

- a) They are somehow similar to subordinated debt, but, unlike bondholders, preferred shareholders could not force a company into bankruptcy if preferred coupons (dividends) were not paid on time
- b) Many preferred shares provide for cumulative preferred dividend payments having priority over ordinary dividends
- c) From an issuer's tax perspective, preferred stocks are a more expensive source of financing than bonds
- d) All of the above

There are generally two kinds of preference shares, cumulative and non-cumulative. The preference share generally pays a fixed dividend that must be paid if an ordinary share dividend is paid. However if a company is doing badly and no dividend is paid the payment on the preference share may not be paid. If it is a cumulative preference stock, this dividend foregone must be repaid if the company recovers, whereas if the preference stock is non-cumulative the dividends foregone are not repaid once the company starts paying dividends. Preference shares rank above ordinary shares for repayment in the event of a company liquidation. The dividends on prefs, like that on the common stock are generally not tax deductible as they are a distribution of shareholders funds rather than debt financing which is generally regarded as an operating expense and therefore tax deductible.

A preferred stock is a hybrid instrument, in which remuneration is variable, but dividends are protected. Generally, if no dividend is paid, the preferred shares become ordinary shares. Assertion c) is relevant for tax purposes only, as preferred shares dividends are not always tax-deductible, contrarily to bond coupons: d).

Convertible bonds

What this is about

Convertible bonds are bonds (fixed income instruments) that can be converted into the equity of the issuing firm at certain times and under certain conditions of the issue. These instruments behave in function of interest rates and the fortunes of the issuing company, providing a floor of value to the investor without depriving him of the upside potential. When the underlying stock price rise the value of the Convertible Bond (CB) will rise as the equity conversion option moves into the money. Conversely, when equity price falls the value of the CB falls back to that of a straight bond issued by that company. Thus CB's are said to give the investor "equity" upside with a "bond floor" and thus have a risk profile between equity and straight bonds.

Readings

Fabozzi Chapters 50 and 51 (pp. 1103 – 1174) (150 min)

Key things you should know

The basic types of CB such as zero coupon, balanced and bond plus warrants, price profile as a function of stock price, how to calculate the conversion premium and conversion price, the embedded options within the CB such as the issuer call option and the investors right to put the bonds back to the issuer if the stock price fails to perform.

Study Questions

Q. A convertible bond is issued at 100% with a USD 10,000 notional value. Each bond gives the right of the holder to convert each bond into 200 shares of the issuing company. If the current share price is USD 40, how much conversion premium is the investor paying for the conversion right?

Each bond costs USD 10,000 and converts into 200 shares that give a cost per share of USD 50. The current share price is USD 40, so the investor is paying a conversion premium of $(50 - 40)/40$ equal to 25%.

Swaps

What this is about

A swap is basically an agreement to lend and borrow money simultaneously to a counterparty but under different interest rate terms, such as fixed rate versus floating rate (can also be differing currency, interest rate, type of redemption, etc.). A huge market has developed in the last twenty years that allows trading of term exposures with a lot of flexibility. Often there is no exchange of principal in the swap and the counterparts agree to pay interest on a notional amount with payments linked to a reference floating rate index such as LIBOR and a fixed swap rate. E.g. A agrees to pay LIBOR plus 10 b.p. per year and receive 5% fixed per year from B for a period of five years. This effectively gives A a risk profile appropriate to a short-term floating-rate loan, whereas B has the risk profile of a long-term fixed-rate loan. Often swaps are used as a very easy way to gain leveraged (geared) exposure to a particular part of the yield curve such as five-year interest rates, when the trader believes they are going to move in his favour. Interest rate swaps are the most common type of swap, but swaps can be done on other assets such as commodities and stock indices. However these types of swaps have slightly differing structures, e.g. in commodity swaps the agreement is to pay a fixed price for an asset against the average price of that commodity (frequently monthly) over the life of the swap.

Readings

Hull Chapter 6 (pp. 125 – 150) (50 min)

Key things you should know

The terminology of fixed and floating-rate, payer and receiver
The reference indices used such as LIBOR
Notional amounts
Exchange of principal at the beginning and end of the swap

Study Questions

Q. A corporate bond is bought in the market at a par yield of 6.50%, the equivalent maturity Government bond has a par yield of 5.75%. If the swap spread for this maturity is 20 b.p. what is the asset swap price of this corporate bond?

An asset swap is a package of a swap and a bond to create a synthetic floating rate note, in this case the spread between the corporate bond and the swap is 55 b.p. since the par swap rate is 20 b.p. above the (Govt.) risk free rate, hence if the corporate bond pays 6.50% fixed with a swap rate of 5.95%, the floating rate payment needs to be LIBOR plus 55 b.p. Hence the asset swap "price" is "LIBOR plus 55".

Caps, floors, swaptions

What this is about

A cap is a series (strip) of call options that have an interest rate as the underlying asset, frequently LIBOR. Each individual option is called a caplet. The floor is a series of put options on interest rates such as LIBOR and each option is called a floorlet. A Cap effectively allows a holder to be protected from rises in interest rates above the cap rate (strike price of the interest rate options) and a floor, in turn, provides protection against falling interest rates. Cap and floors are generally used to protect the amount of interest to be paid or received on floating rate loans.

Swaptions are options on interest rate swaps – they are either payer or receiver swaptions that give the right (but not the obligation) to pay or receive the fixed leg in the underlying swap. As an example a European style swaption may be referred to as a 1y into 5y payer swaption that would allow its holder the right to chose to pay fixed (and receive floating) on a 5-year swap at the end of a one-year option life.

Caps and floors are instruments to trade long-term interest rate exposure, in that they provide, at regular times, the difference (or nothing) between an interest rate and an agreed level. Swaptions are options on swaps.

Readings

Hull Chapter 22 (pp. 508 – 536) (60 min)

Key things you should know

Basic structure of Caps and floors in terms of component options, understanding of average strike rates for Caps and Floors, understanding swaption terminology, payer and receiver and relationship between the term of the swaption and the underlying swap.

Study Questions

Q. A Cap consists of three caplets that expire in three months, six months and nine months. The holder of the cap is paying 3m LIBOR set in advance and paid in arrears on a floating rate loan that will be repaid in nine months time. The cap

rate is 5% and the loan amount is USD 10m. If the LIBOR fixing today is 6% for payment in three months time, what is the approximate value of the first caplet at exercise?

The LIBOR rate on the loan is fixed today and payable in three months time, this rate is 6% for three months on USD 10m. But the caplet has a strike price of 5% on the same notional, so the option payoff will protect the loan holder from the higher interest rate. Without the cap, he would pay $(6\%/4) \times \text{USD } 10\text{m}$ but with the cap he effectively pays $(5\%/4) \times \text{USD } 10\text{m}$ so the caplet must pay out USD 25,000 to compensate for the higher rate. Obviously the overall cost to him also includes the premium paid for the cap and this will effectively give him a marginally higher cost as the price for buying protection.

Simple exotics

These include barrier options.

What this is about

Simple or vanilla options such as Calls and Puts allow the right to buy or sell the underlying asset at a fixed price. Exotic options have more complex pay off conditions that modify the risk return profile. Barrier options, for example, can be options that are cancelled (knocked out) at certain levels of the asset price. The risk profiles, hedging and pricing of exotics can be complex. There are various classes of exotic including barrier/ trigger types, frequently called "path dependent" options, cross currency or currency protected options, multi asset options based on more than one underlying asset. Theoretically there is no limit to the variations of exotic that can exist (e.g. consider a pay-off function of $(\text{Spot} - \text{Strike Price})$ raised to the power n , where n can be any integer!).

Options and futures have been developing beyond the plain vanilla definition, offering a range of conditions for payments to occur, based upon market prices in the future.

Readings

Hull Chapter 19 (pp. 435 – 455) (40 min)

Key things you should know

Definitions of simple exotics

Barriers, averages, currency-protected, basket, best of, difference of, options

Pay-off function of each exotic

Qualitative assessments of risk, e.g. barrier option where the knock-out level is above the level where the option has intrinsic value.

Study Questions

Q. An up-and-out knock out call option has a strike price of 100 and a knock out level of 110. What are the major risk issues in trying to delta hedge this option?

With the spot price below 100 the option behaves much like a normal call option with a low delta rising as the option moves in to the money. However when the option approaches the barrier the option has intrinsic value close to 10 units, but once it crosses the barrier it has zero value. Hence the delta hedge must have the opposite risk profile and increase in value by almost 10 units when the spot price moves through 110. This means that the delta hedge must be many times the

underlying value (notional) of the single option and hence suffer from liquidity problems, i.e. the hedging process itself causes the knock out of the option and the market moves wildly against the hedger causing excessive hedging costs.

Markets

This part includes 40% of the points of exam module I, and therefore 10% of the marks for the entire PRM qualification examination.

Money market / FX market

What this is about

Money markets are where financial operators such as banks can lend and borrow short-term money, typically deposits (short term loans) for periods of up to one year. Frequently, these short-term deposits are represented by bills or other instruments such as commercial paper (in effect company IOUs) that are traded directly. FX (Foreign Exchange, or FOREX or FX) markets are where to buy and sell foreign currencies. Trades involve either spot trades (for today or very near future) or forward trades for settlement in up to one year's time. In these markets, the most standardised and liquid short-term instruments can be traded.

Readings

Bodie

Skim chapter 1 for 'culture' (this should take no more than 15 min), and chapter 2 section 1 in more depth (20 min)

Reuters' FX MM, pp. 7-321 (600 min)

Key things you should know

The role of the international money markets

Liquidity and tradability

The interbank deposit market

OTC and traded instruments

Issuers of money-market instruments

The different money market instruments: Bills, CDs, CPs, FRAs, and repos

Yields of money-market instruments

Type of quotation

Spot and forward currencies

FX options and futures, FX swaps

The Euro

Study Questions

Q. Two otherwise identical instruments are quoted in different currencies, both freely and widely traded. What would be the main factors to explain price differences?

- a) The countries' credit ratings
- b) The relative interest rates in both currencies
- c) The liquidity of both currencies
- d) The correlation of both currencies to USD

Whereas a) and c) would intervene to a small extent (say, some 10s of basis points), d) would not, b) would be most relevant, as differences in interest rates between currencies are quite significant (commonly some 2-5% within OECD): b)

Capital markets

What this is about

Capital markets buy and sell financial instruments that provide finance for the world's corporations and financial institutions. These instruments include bonds, equity, Convertible Bonds and many variants. Primarily the banks issue financial instruments on behalf of corporates to raise money, which in turn, are bought by investors as a way of employing their surplus capital and earn a return.

Capital markets include short-term money markets and markets for long-term capital, together with the primary financial instruments and derivatives commonly used for long-term operations. They include fixed-income and equity markets. Beyond the most common derivatives, there are lots of derivative products designed to provide unique risk-return characteristics that are private (over-the-counter or OTC) transactions between two market counterparties.

Readings

Bodie

Chapters 2.3 to 2.5 (60 min), 3 (20 min), 20 (30 min) and 22 (30 min)

Hull

Chapters 1 and 2 (together 50 min), 5 (30 min) and 7 (30 min)

Key things you should know

The role of financial intermediation

Capital, investment and risk

Interest rates, yield curve, zero curve

Risk premium

Arbitrage

Fixed-income and equity-based markets, derivatives

Options and futures

Common equity and preferred stock

Stock Exchange indices

Equity valuation

Study Questions

Q. Put the following instruments in order of what would capital markets offer in terms of expected returns:

- a) Equity of a well-established chemistry company
- b) Foreign (OECD) government long-term bond
- c) CD issued by a BB bank
- d) Interbank deposit

The instruments above present different levels of risks, and therefore different levels of expected returns. The most risky instrument is the equity, although it is a chemical company, generally less volatile. A long-term bond would have credit risk

and currency risk. A BB-rated bank would not be as secure as a deposit on the interbank market. Hence: a-c-b-d, in decreasing order.

Markets for commodities

In general, knowledge of the global markets for natural gas and oil (energy) will be tested. Commodity markets are split roughly between soft commodities such as sugar and cocoa, metals and ores such as copper and energy such as oil and gas. The markets are further split between raw materials or feed-stocks and products (e.g. Brent crude and unleaded gasoline). Commodity markets have special characteristics over and above financial markets, these include:

- Finite supply
- Costs of extraction
- Costs of transportation
- Costs of storage
- "Shelf life"
- Advances in technology

These markets therefore have highly individual characteristics that require extensive specialist knowledge for successful risk management.

What this is about

A huge market exists for these commodities. These can be traded spot or forward, or via commodity swaps. These transactions can be for "physical" or paper delivery. These markets generate price volatility through the effect of short-term physical supply and demand, but also, and in the very short-term more importantly, the price expectations of physical operators and speculators. The extent of the possibility to store commodities, and the geopolitical situation, give markets for commodities much of their idiosyncratic behaviour.

Readings

Reuters' Commodities, Energy and Transport, John Wiley & Sons Finance, pp. 7-60 and pp. 257-288 (160 min)

Key things you should know

Spot, forward trading
Exchange-traded contracts
Expected prices, future prices
Contango, backwardation
Storage and delivery
Volatility, price peaks

Study Questions

Q. What could be an explanation for the market for oil being in backwardation?

Backwardation in a commodities market means that the spot price of the commodity is higher than the price for forward delivery, taking into account storage and other costs of carry. It is usually brought about by a short term increase in spot demand which is expected to subside shortly (for instance a sudden cold snap in winter driving up short term heating oil prices) and can also occur when this increase in short term demand occurs in a market which needs time to respond to increased demand by increased supply, e.g. there is a delay of

several weeks / months between crude oil being pumped in the Middle East and gasoline being purchased at a filling station in the Mid-Western United States. Seasonal factors affecting demand could be a factor. Operators expect the price of oil to drop within a certain time frame. The political situation in a major oil-producing region could be a factor, as well as the expectation that operators release inventories held for precaution. The nature of these anticipations by market operators is the key factor.

Power markets

What this is about

Power is different from other commodities as, in the form of electrical power, it is not directly or easily storable, and therefore spot/forward prices are more volatile, and possibilities for transportation are more limited, hence these markets are more fragmented. These aspects, together with the presence of particular regulation specific to domestic energy supplies, make power markets less easy to operate in than capital markets.

Readings

Reuters Commodities, pp. 63- 130 (120 min)

Key things you should know

Why power markets are different
Contract terms in power markets
Convenience yield
Price determinants in power markets
Forward price curve
Volatility in energy markets

Study Questions

Q. What are the main impediments to successfully predicting power market prices with random walk theory?

These are numerous, starting with:

- seasonality of prices, intra-day, within the year, etc., hence prices tend to follow sinusoids (sine waves, or regular up-and-down cycles), rather than straight lines, before the error terms are added
- presence of shocks (power disruptions and others), breaking trend lines
- error terms, due to numerous market frictions e.g. weather and difficulties of delivery, are larger than in financial markets

Exam II

In this exam, knowledge of the mathematical foundations of Risk Measurement and pricing will be tested. It lasts 2 hours and counts for 20% of the total exam points. The total time estimated to read the source texts for this module is 50 hours (competent mathematicians will require much less).

The mathematics needed for the exam go beyond what is needed in finance day-to-day, but are generally at the basic university level of mathematics. This is to

insure that successful candidates are equipped for the quantitative challenges of tomorrow.

Calculus

This part counts for 25% of the exam II module and hence 5% of the whole PRM exam marks

The calculus of variations deals with modelling rates of change and includes the differential and integral calculus. Differential and integral equations can be shown in general to be equivalent. Differential equations can characterise market movements and are used to model derivatives prices via equations like Black-Scholes.

This is critical to determine, in a risk environment, the effects of changes in the environment on the fortunes of the institution. No risk manager will survive if not calculus-literate.

Readings for this part include:

Calculus

Very basic preparation: Chapter 1 – 8 (pp. 1 – 78) (240 min)

Basic preparation: Chapter 10 – Chapter 22 (pp. 86 – 205) (360 min). For the candidates with a strong mathematics background, much of this section can probably be omitted. Have a look at the questions as a guide.

Rate of change

What this is about

The basic concept of how a function value changes as the independent variable changes – i.e. the slope or gradient of curves.

This is a simple algebraic tool to determine how outputs change in relationship with inputs. This tool must be well mastered to understand further stages of calculus.

Readings

Calculus Chapter 9 – pp. 79 – 86 (25 min)

Key things you should know

Definition of rate of changes

Definition in the limit

Derivative notation

Differentiability

Study Questions

Q. Find the derivative of $y=2x$ using rate of change approach:

$$\Delta y = f(x+\Delta x) - f(x) = 2(x+\Delta x) - 2x$$

$$\Delta y = 2\Delta x$$

$$\Delta y / \Delta x = 2$$

$$\text{In the limit } dy/dx = 2$$

Area/volume

What this is about

The area and volume of simple geometric figures can be calculated with simple formula learned by every of us in early stages of schooling. The main challenge is not to have forgotten about these. The area under the function is calculated using the integral of the function (elemental area) between the appropriate limits. Volumes of revolution are calculated by integrating over the elemental volume.

Readings

Calculus Chapter 23 (pp. 206 – 215) (30 min)

Calculus Chapter 29 – 30 (pp. 257 – 280) (70 min)

Key things you should know

How to evaluate standard integrals

How to calculate volumes and areas for simple functions

Study Questions

Q. Calculate the area under the curve $y = x^2$ for the range zero to one

$$\text{Area} = \int_0^1 y dx = \int_0^1 x^2 dx = \left[\frac{x^3}{3} \right]_0^1 = \frac{1}{3}$$

Optimization

What this is about

Given a numerical relationship, or a function, optimization tools allow determination of the appropriate input to maximise/minimise output.

The simplest approach looks at function maxima and minima using the first and second derivatives. For more complex problems partial derivatives and Lagrange multiplier coefficients can be used to that same effect.

Readings

Calculus Chapter 14 (pp. 115 – 128) (40 min)

Key things you should know

How to calculate maxima and minima for simple functions

Study Questions

See worked examples in chapter 14

Taylor series expansions

What this is about

When functions are not easily traceable, this is a handy tool to replace these, at a given point, by a polynomial approximation, called the Taylor series expansion. The Taylor series represents the function value by a polynomial with coefficients in ascending orders of the derivative of the function evaluated at that point.

Readings

Calculus Chapter 47 (pp. 432 – 441) (30 min)

Key things you should know

How to calculate a Taylor series expansion for a simple function

Study Questions

See worked examples in chapter 47

Ordinary and partial derivatives

What this is about

The rate of change in functions given the change of the input values is defined as the derivative. This tool is key to determine what input is critical in a numerical relationship, and what will happen to the output if an input changes by a certain amount. Ordinary derivatives look at rates of change of the function dependent on all independent variable simultaneously; partial derivatives gauge the rate of change of function values on one independent variable whilst others are held constant.

Readings

Calculus Chapter 48 (pp. 442 – 451) (30min)

Calculus Chapter 49 (pp. 452 – 463) (30 min)

Calculus Chapter 59 (pp. 559 – 572) (40 min)

Key things you should know

How to differentiate simple functions, the difference between full and partial derivatives, solving simple differential equations by separation of variables and integration.

Study Questions

Q. Find a linear polynomial $p(x)$ that is a tangent-line approximation for the function:

$$f(x) = e^{2x - 4}$$

at the point $x_0 = 3$.

- a) $14.778x - 36.945$
- b) $7.389x + 2.718$
- c) $2.718x$
- d) $14.778x + 0.018$

The tangent to this function will pass through the same value as the function at $x=3$ and will also have the same gradient. So we use the first derivative of the function to establish the slope of the tangent and use the normal $y=mx+c$ representation of a straight line.

$2x-4$, at the point $x_0=3$, is worth 2, 2.718^2 is worth 7.39. The tangent at the point 3, is sloping at $2 * e^{(2*3-4)} = 14.778$: this discards a) and c). The d) line, at the point 3, is 44.35: too high, only a) remains. To double-check, $14.778 * 3 - 36.945 = 7.39$: a)

Integration

What this is about

This calculates the area below/above the curve of a given function, or the functions of which derivatives gives back the original function (i.e., integration as the reverse process to differentiation). This tool is fundamental to evaluate cumulative effects of phenomena.

Readings

Calculus Chapter 31 (pp.: 281 – 288) (25 min)

Calculus Chapter 32 (pp.: 289 – 304) (45 min)

Key things you should know

Integration by parts, standard substitutions and standard form integrals.

Study Questions

Q. Evaluate the definite integral:

$$\int_0^2 x e^{x^2} dx$$

- a) 5.437
- b) 26.799
- c) 21.285
- d) 7.389

This integral can be done by noting that e^{x^2} differentiates to (using the chain rule)

$2xe^{x^2}$ so the integral $I = \frac{e^{x^2}}{2}$ which on putting in the limits $I = \left[\frac{e^4}{2} - \frac{1}{2} \right] = 26.799$,

so answer b)

Linear algebra

This part counts for 25% of the exam II points and hence represents 5% of the total marks for the PRM exam

Linear algebra is used in every-day situations in a risk management environment. The use of VaR has made it more critical than before to use vector calculus, and matrix calculus.

Please note that basic one-dimensional algebra is not mentioned in the list of topics. This is a prerequisite for this part of the exam, as this is the basis for matrix calculus.

Matrix algebra

This includes determinants and singular matrices.

What this is about

Matrices are a structured representation of numbers for assessing the effects of multidimensional relationships. They allow manipulation of multidimensional data in a structured and controlled manner. Matrices are an indispensable tool for dealing with variance-covariance analysis and VaR. Matrices have special properties such that, in general, the order of multiplication matters so that AB is not equal to BA , they are said to be non-commutative. The determinant of the matrix is a special product of its component elements and the determinant has to be non-zero for the matrix's inverse to exist. A singular matrix has no inverse hence its determinant must be zero.

Readings

"Linear Algebra", Schaum's

Chapter 1 pp. 1 - 23 including all worked examples, do a selection of five supplementary questions (90 min)

Chapter 2 pp. 28 - 53 including all worked examples, do a selection of five supplementary questions (60 min)

Chapter 3 pp. 59 - 111 including all worked examples (160 min)

Chapter 4 pp. 116 - 162 including all worked examples (140 min)

Chapter 8 pp. 277 - 301 including all worked examples (80 min)

Chapter 9 pp. 306 - 335 including all worked examples (90 min)

Key things you should know

Matrix and Vector representations

Adding and subtracting matrices and vectors

Scalar, dot, cross and vector products

Matrix and Vector algebra

Solving systems of linear equations

Vector spaces

Determinants

Diagonalisation of matrices

Determination of eigenvalues and eigenvectors

Study Questions

Q. Which of the following statements is false:

- Singular matrices have determinant 0.
- Singular matrices have columns that are not independent vectors.
- A product of two non-singular matrices can be singular.
- Singular matrices have 0 as an eigenvalue.

It is enough to find a statement that 'looks' bizarre and then to read it more thoroughly to refute it. However, if no inspiration comes at first reading: a non-singular matrix is defined as having a non-zero determinant (a) is correct); a zero determinant is an indicator of dependence of vectors (b) is correct). If the determinants of two matrices is non-zero, their product cannot be 0 (c) is wrong). As for d), it is equivalent to a), hence correct: answer is c).

Q. Determine the inverse matrix of:

$$\begin{pmatrix} 1 & 2 \\ 0 & 2 \end{pmatrix}$$

- a) $\begin{pmatrix} 1 & 1 \\ 0 & 0.5 \end{pmatrix}$
 b) $\begin{pmatrix} 1 & -1 \\ 0 & 0.5 \end{pmatrix}$
 c) $\begin{pmatrix} 0.5 & -1 \\ 0 & 1 \end{pmatrix}$
 d) $\begin{pmatrix} 1 & -1 \\ 0.5 & 0 \end{pmatrix}$

It is quicker to eliminate the matrices by multiplying them by the first matrix, as the product needs to be the identity matrix if it is the inverse. We can eliminate a), as it has no negative element. The matrix b) works. The other two matrices would fail to produce the desired Id matrix (as c) fails on first line first column, d) second line first column): b)

Positive definiteness

What this is about

A real symmetric matrix B is positive definite if the product of a transposed non-zero vector U^T times B times U is greater than zero. If B is positive definite, the result is an inner (or scalar) product.

The determination of VaR often involves the use of matrices containing market data, the solution requires the manipulation of this data using matrix techniques were the result needs to be positive and real (as it is a real world risk number) – hence it is important the data matrix is positive definite prior to the calculation.

Readings

Linear Algebra Chapter 7 page 248 (bottom) – 249 (4 min) Chapter 13 p. 400 (2 min)

Key things you should know

The definition of positive definite, what are the implications if a matrix is positive definite for matrix algebra.

Study Questions

Q. Under what circumstances is the 2 x 2 real symmetric matrix, A, positive definite?

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

In order for A to be positive definite, a and b both have to be positive and the determinate of the matrix must be positive.

Eigenvectors and eigenvalues

What this is about

If any square matrix (number of rows = number of columns) can be multiplied by a non zero column vector such that the result is a scalar (number) times that column vector, then vector is known as the eigenvector and the scalar is the eigenvalue of the square matrix e.g. $A*u = L*u$ where A is a square matrix and L is a pure scalar. Eigenvalue equations are characteristic of matrices which can be diagonalised, and diagonalised matrices are relatively easy to manipulate to get a solution e.g. in the case of VaR calculations. Diagonalisation of the matrix is equivalent to factorisation into the component basis.

Readings

Linear Algebra Chapter 9 pp.: 310 –316 (20 min)

Key things you should know

Definition of eigenvalue equation

Relationship between eigenvector equations and the diagonalisation (factorisation) of square matrices

Calculation of eigenvalues and vectors for two-by-two matrices.

Study Questions

Q. Show that the following vectors v_1 , and v_2 are eigenvectors of A – what are the eigenvalues?

$$v_1 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}, v_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, A = \begin{pmatrix} 3 & 1 \\ 2 & 2 \end{pmatrix}$$

$$A * v_1 = \begin{pmatrix} 1 \\ -2 \end{pmatrix} = v_1, \text{eigenvalue} = 1$$

$$A * v_2 = \begin{pmatrix} 4 \\ 4 \end{pmatrix} = 4 * v_2, \text{eigenvalue} = 4$$

$$v_1 \neq k * v_2$$

The eigenvectors are linearly independent.

Cholesky factorisation

What this is about

If a real symmetric matrix can be split into two factors such that it is equal to the product of these two factors, the first of which is a lower triangular matrix (with zeros in the upper right hand area) and the second is the transpose matrix with zeros in the bottom left area, this is termed a Cholesky factorisation. This factorisation is used in determining VaR when the correlation matrix is factorised into its two components. This decomposition of the correlation matrix will fail if the number of independent risk factors is less than the dimensionality of the matrix; hence this decomposition is useful if we want to ensure that the risk factors used are truly independent. Be aware that there is also the Cholesky decomposition that is used to sample multivariate distributions.

Readings

Jorion Chapter 12 Section 3.2 (20 min)

Key things you should know

How to perform the factorisation
 Uses in VaR
 Significance if it does not exist

Study Questions

Q. Perform the Cholesky factorisation on the following correlation matrix:

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} = \begin{pmatrix} n_{11} & 0 \\ n_{12} & n_{22} \end{pmatrix} * \begin{pmatrix} n_{11} & n_{12} \\ 0 & n_{22} \end{pmatrix} = \begin{pmatrix} n_{11}^2 & n_{11}n_{12} \\ n_{11}n_{12} & n_{12}^2 + n_{22}^2 \end{pmatrix}$$

By equating elements of the matrix and eliminating terms, we get:

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \rho & (1-\rho^2)^{1/2} \end{pmatrix} * \begin{pmatrix} 1 & \rho \\ 0 & (1-\rho^2)^{1/2} \end{pmatrix}$$

Probability and statistics

This part counts for 50% of the exam II module and hence 10% for the whole PRM exam marks.

Evaluation of future outcomes generally involves the analysis of the past. Using statistical measures allows the characteristic nature of large volumes data to be analysed. Probability measures allow estimates of likelihood of events based on past observations. Probability and statistics are widely used to characterise and forecast risk and to analyse data to search for causal relationships.

To understand uncertainty, the best that has been found so far is to associate to a possible future event a number relating to its likelihood: in other words, probabilities. This is the main, if not the only way, to characterise uncertainty, and hence risk.

Probabilities can be seen as related to one single event, or to a set of events that can occur separately, in other words to multiple risks. Therefore, the notion of correlation is introduced, to relate to diversification. Diversification is a fundamental tool to protect against risk. The effectiveness of diversification can be measure through the notion of correlation.

Readings for this part of the exam include Schaum's Outline Probability and Statistics:

Chapter 3 (pp. 78 – 112) (100 min)
Chapter 5 (pp. 161 – 204) (130 min)
Chapter 7 (pp. 224 – 277) (160 min)
Chapter 8 (pp. 278 – 327) (150 min)
Chapter 10 (pp. 363 – 388) (75 min)

Random variables

What this is about

These are variables that have a random nature such that individual observations of the variable may be impossible to forecast, but average characteristics such as standard deviation or mean of the distribution from which the random variable is drawn may be estimated. Examples of random variable might include e.g. stock price behaviour in a simulation of hedging an option portfolio.

Readings

Probability and Statistics, Chapter 2 (pp. 36 – 77) (120 min)

Key things you should know

Properties of random variables (finite and infinite populations)
Algebraic properties
Common probability distributions such as Normal (Gaussian) and Poisson

Study Questions

Q. What can we say about the sum $X + Y$ of two independent normal random variables X and Y :

- a) It is normal only if X and Y have the same mean.
- b) It is always normal.
- c) It is chi-squared.
- d) It is chi-squared if X and Y both have mean 0.

One can refute c) and d) by taking a normal distribution with a zero standard deviation (it is just a number): add this to a normal distribution, and it will give a normal distribution. Adding two normal 0-variance distributions with different means, will give a normal distribution (with 0 variance, too), discarding a). A more elegant resolution is to remember the statistics course, to recall that is the sum of two normal independent distributions is itself a normal distribution, and go immediately to answer b).

Distributions and densities

This includes the general theory for univariate probability densities plus knowledge of standard distributions (uniform, normal, lognormal, Poisson, Chi-squared etc., conditional probabilities).

What this is about

The probability density function defines how the probability of an event is distributed over an interval. The integral over the distribution should equal one, as this corresponds to the sum of the probabilities of all possible outcomes. The density function allows the calculation of a event occurring in a interval of the random variable, so for example we could estimate that the probability of a stock price rising by less than one percent tomorrow (based on history) is 66% where the stock price is the random variable and we choose a standard distribution function (e.g. lognormal) as the representative distribution. The most common distribution is the normal distribution and it can be shown that in the limit of a large number of uncorrelated distributions (of any kind) the net result will be normal distribution. Some distributions are continuous (i.e. defined for any value of the variable) where some are discrete (e.g. Poisson) which is defined only for discrete values e.g. 1,2,3,4 etc.

Readings

Probability and Statistics, Chapter 2 (pp. 36 – 77) (120 min)

Probability and Statistics, Chapter 4 (pp. 113 – 158) (135 min) (special distributions)

Key things you should know

How to calculate the moments of distributions, definitions and properties of the named distributions, how to integrate the density functions over an interval to get probability of an event occurring in a range, basic applications of each type of distribution.

Study Questions

Q. What is the standard deviation of a random variable Q with probability function

$$\phi(q) = \begin{cases} .25 & q = 0 \\ .25 & q = 1 \\ .50 & q = 2 \end{cases}$$

- a) .6875
- b) .4727
- c) .8291
- d) .4281

We assume in the question that the distribution is discrete – so we calculate a simple mean equal to $0*0.25+1*0.25+0.5*2=1.25$

Variance= $((1.25)^2)*0.25+((0.25)^2)*0.25+((0.75)^2)*0.5$

STD= $\text{SQRT}(0.6875)=0.8291$

So answer c)

Moments

The moments up to the fourth one will be considered.

What this is about

The moments of a distribution are measures of its characteristic shape and general properties. The simplest moment is the mean, then the variance, then skewness and the fourth moment, kurtosis. These moments are determined by computing the sum of the differences between each data point in the distribution and the mean raised to successively higher powers, so variance for example is the sum of the differences between the mean and the data point raised to the power 2, and is hence call the second moment of the distribution.

Readings

Chapter 3 (pp. 78 -112) (100 min)

Key things you should know

The definitions of the moments

How to calculate moments for continuous and discrete distribution

Interpretations of moments e.g. kurtosis, leptokurtic

Study Questions

Q. What is the formula for the skewness of a random variable X that has mean μ and standard deviation σ ?

a) $\frac{E([X - \sigma]^2)}{\mu^2}$

b) $\frac{E([X - \mu]^4)}{\sigma^4}$

c) $\frac{E([X - \mu]^3)}{\sigma^3}$

d) $\frac{E([X - \mu]^4)}{E([X - \sigma]^4)}$

The solution d) can be eliminated as it makes little sense to use standard deviation in a sum for the divisor. It is worth remembering that skewness is the 3rd moment of a distribution. Above, a) looks like variance but is divided by the mean squared, b) is the 4th moment (kurtosis): so the skewness is answer c)

Q. Your general manager tells you with some anxiety that, out of the last calendar year, although VaR has been decreasing while the bank has been more active in the markets, as much as 51% or so of the daily VaR figures are above the median value for the year. What do you do?

a) You recommend a reduction of VaR

b) You find a quick explanation for this, and do nothing else

- c) You recommend a relaxation of trading limits
- d) You check your figures

Better remember what is a median (namely, the value for which half the population is above, half is below), so if 50% give-or-take of the values are above the median, this is nothing else than the definition of the median, so answer b)

Covariance and correlation matrices

What this is about

These matrices measure the amount of correlation (moving together) of variables and are the basis of VaR calculations. Covariance and correlation matrices are equivalent – they differ only in the fact that one is normalised by division by the variances.

Readings

Linear Algebra Chapter 3 (pp. 96 - 97) (10 min)

Key things you should know

The definitions of covariance and correlation
Characteristics of matrices from the Linear Algebra
Use in the VaR calculation from Jorion Chapter 7 [add?]

Study Questions

Q. A covariance matrix for a random vector:

- a) Is strictly positive definite, if it exist
- b) Is non-singular, if it exist
- b) c) Always exists
- d) None of the above

This question is full of red herrings. The main thing is not to get bogged down into what is a random vector, and to find out that the only relevant fact is that a covariance, as the name indicates, is between two objects. Hence d)

Principal component analysis

What this is about

PCA is a method of analysing data to model its behaviour in terms of a small set of characteristics that explain most of the behaviour of the data. For example yield curve data can be analysed over time and modelled in terms of two principal components, for example parallel shifts up and down and rotations of the short-term rates versus the long-term rates. These two model factors “explain” most of the data structure in terms of two principal components. PCA is therefore a method of mining data to look for causal relationships, these are very useful for risk management, as they allow more efficient hedging.

Readings

Hull Chapter 16 (pp. 360 – 364) (15 min)

Key things you should know

The basics of PCA, why it is useful, how it might be used in VaR

Study Questions

Q. Why is PCA useful for risk management?

PCA allows the hedging to be carried out with a reduced number of hedge instruments as it allows the “normal” models of the risk to be identified and hedged directly rather than using bucketing or position-by-position approach.

Monte Carlo simulation

What this is about

Monte-Carlo allows the simulation of random behaviour according to given parameters (e.g. mean, standard deviation) to evaluate real world outcomes by simulating the underlying process. For example MC can simulate stock price movements and the efficiency of a hedging scheme can be evaluated by running the simulation many times. It can also be used to randomly sample values of multidimensional functions for purposes like numerical integration of multidimensional problems (e.g. best of n asset options).

Readings

Hull Chapter 18 (pp. 410 – 414) (15 min)

Key things you should know

How MC can be used for random sampling and for the simulation of random processes. The slow convergence for MC that is order \sqrt{N} where N is the number of simulations for each problem.

Study Questions

Q. How can a random number generating function be used to generate samples from a normal distribution?

By using a sum of a large number of independent random numbers from a uniform distribution such as is generated by a random number function, it is possible to approximate a normal distribution. Usually twelve samples is considered large enough, so our normal random variable is the sum of twelve random numbers minus the mean (6).

Linear regression

What this is about

This is concerned with fitting data so that a characteristic straight line can be calculated which minimises the distance, on average, between the line and the data points. The characteristics of the line (i.e. slope and intercept) allow a simplified view of the relationship between factors e.g. a single equity price and

the movement of the market as a whole. The significance of this relationship is related to the "R squared" regression coefficient.

Readings

Probability and Statistics, Chapter 8 (pp.: 278 –289) (35 min)

Key things you should know

How to calculate a least-squares fit
Regression coefficients
Hypothesis testing

Study Questions

See Probability and Statistics Chapter 8 Q8.2 p. 289

Basic statistical tests

What this is about

These allow the statistical testing of hypothesis about data sets, for example, is a small finite data set actually derived from a normally distributed process? Or are two sets of data consistent with the same underlying distribution? Most common tests are student-t, chi-squared.

Readings

Probability and Statistics Chapter 7 (pp. 224 – 277) (150 min)

Key things you should know

How to carry out basic tests using student-t and chi-squared
Levels of significance
Hypothesis acceptance or rejection

Study Questions

See Probability and Statistics Chapter 7 p. 243 Q7.16

Coping with missing data

What this is about

This concerns dealing with sparse information and how one can use a small data set to infer further information. For example fitting a polynomial through data points in order to calculate intermediate values, this is frequently used to construct yield curves and discount functions were a relatively small number of government bonds define data points and a series cubic functions are fitted through the points (cubic splines). Interpolation generates values between data points, extrapolation generates values beyond the data set.

Readings

Probability and Statistics Chapter 5 (pp. 161 – 203) (120 min)

Key things you should know

How to use sample data to estimate larger populations with magnitude of possible error, curve fitting, interpolation and extrapolation functions

Study Questions

Q. You are given a small series of discount government bond prices and asked to construct a yield curve for valuing a swap portfolio, what method would you use to cope with the sparse / missing data?

The curve needs to be defined for each day as a cash flow may occur on any day, therefore we need a function which is defined between data points, a series of cubic polynomials with knot points such that the curve value and first derivatives are equal will provide a smooth continuous function which closely follows the expected shape of the yield curve without creating unfeasible kinks or deviations.

Exam III

In this exam, knowledge of Risk Management practices will be tested. It lasts 90 minutes and counts for 30% of the total exam points. The total time estimated to read the source texts for this module is 35 hours.

Market Risk

This part counts for 33% of the exam III module, and hence 10% of the total marks of the whole PRM exam.

Duration and convexity

What this is about

Duration and convexity are measures of sensitivity of a bond price to changes in yield. Duration is the first derivative and convexity is the second derivative, hence convexity is the rate of change of duration with changes in yield. There are several different definitions of duration, which differ by factors like $(1+\text{yield})$. Zero-coupon long-dated bonds have the highest duration. These indicators are critical to determine the behaviour of a given bond, and are commonly used in managing the risk of a bond portfolio.

Readings

Hull Chapter 5, sections 5.13 and 5.14 (60 min)

Fabozzi, chapter 5, pp. 99-123 (60 min)

Crouhy, chapter 5 appendix 1 (30 min)

Key things you should know

Calculation of duration

Interpretations of duration

Mac Aulay, modified duration

Graphical representation
Convexity adjustment

Study Questions

Q. When does the duration of a bond equal its maturity?

Duration being the weighted average of the maturity of (discounted) cash flows, duration and maturity equate when there is only a final cash flow to be received at maturity (in other words, for a single cash flow, or a zero-coupon bond).

Cash flow maps, PVBP and interest rate sensitivity

What this is about

Fixed-income instruments can, for portfolio analysis purposes, be mapped into a small finite set of zero-coupon bonds in a consistent way, and then regrouped by maturity. This is called mapping cash flows. PVBP is the change in present value of the portfolio for a 1 basis point change in yield, which is the interest rate sensitivity for a 1 b.p. move.

Readings

Hull, appendix 16A, p. 368-369 (20 min)
Fabozzi, chapter 5, p. 124-127 (20 min)

Key things you should know

Practical use of PVBP
Yield and price volatility
The cash-flow mapping process

Study Questions

Q. What happens to the variance of a bond portfolio when cash-flow mapped?

The aim of such procedure is to aggregate the constituents of a portfolio into a set of cash flows with different maturities, to be used for purposes of sensitivity analysis. The variance is preserved as the same, as a change would invalidate risk analysis.

Greeks of instruments and portfolios

What this is about

All instruments within a portfolio will react in different ways to changes in environmental variables such as discount rate. 'Greeks', or sensitivities, help to perform portfolio-wide analysis of the effect of a change in a variable on the whole portfolio. The most common Greeks are the Delta, Gamma, Vega and Theta of the position.

Readings

Hull chapter 14 (40 min)

Key things you should know

Greeks: definition, methods of calculation
How to hedge a portfolio for different variables
Scenario analysis
Portfolio insurance

Study Questions

Q. A portfolio of bond options is delta-hedged. What risks is it still exposed to?

The portfolio is still exposed to the other 'Greek' risks: vega, rho, and theta. If the underlying asset moves in price, the delta of the portfolio will change, making the original delta-hedge either insufficient or excessive: this is the gamma risk.

Implied volatility and smile

'Smirk' is within the syllabus. This concept would represent a non-symmetric variation of smile.

What this is about

A reverse calculation from the market price of an option into the volatility of the underlying asset, using an option model like Black-Scholes produces market-implied volatilities. These vary not only with the maturity of the option (term structure), but also with the distance of the underlying to the strike price. The shape of this implied volatility curve can have a shape like a "smile" or a "smirk" so that the implied volatility of the out of the money options is higher than the implied volatility of the at the money options.

Readings

Hull chapter 15 (30 min)

Key things you should know

Fat tails
Implied volatility calculations
Volatility smile
Volatility surface

Study Questions

Q. Does the volatility smile come from theory or market practice?

The commonly used option pricing models assume normal returns, hence log-normal distribution of prices, with no account for kurtosis. Practice and empirical studies, show that returns are distributed differently from the assumed normal distributions, with 'fat tails'. Hence the probability of a far out-of-the-money option to be exercised is higher than suggested by a model using flat volatilities. Hence the volatility smile is a market practice solution to an incomplete model.

Value-at-Risk (VaR)

What this is about

Value at Risk (VaR) is a fundamental tool that helps determine, within a specified confidence interval, how much money could be lost on a portfolio over a given period of time. For example, the 1-day 99% confidence interval VaR may be USD 10m, which means that we expect to lose USD 10m or more (and it could be much more!) on only 1 day in 100. VaR takes into account the diversification effect. A considerable amount of data and assumptions are needed to integrate all market risks within a portfolio. The most common approaches to evaluating input data are parametric and historical approaches. The simplest method is probably the delta-normal approach, which uses the portfolio delta and the standard deviation of the risk factor assuming a normal distribution. Generally speaking, the largest problem in large VaR calculations is handling enormous amounts of data and obtaining long market histories.

Readings

Jorion chapters 3 (Section 3.3), 5-11, 14 and 21 (Section 21.1) (200 min)

Hull chapter 16 (60 min)

Linsmeier (100 min)

Crouhy chapters 1 and 5 (section 1) (120 min)

Key things you should know

Risk Management systems

VaR as a tool to manage risk

VaR as a regulatory requirement

Study Questions

Q. How does VaR change if the holding period goes from 1 to 10 days, under the usual set of assumptions?

Usual assumptions include normal returns and no autocorrelation, hence the variance of returns will be proportional to the number of days. The standard deviation, hence the VaR, will be proportional to the square root of the number of days, hence will be multiplied by $\sqrt{10} = 3.16$.

Calculation of VaR for linear portfolios

What this is about

The calculation of VaR entails a good grasping of the individual components of a portfolio, including the statistical models to apply to future cash flows and to exposures. Linear portfolios have delta but a zero gamma, hence the risk sensitivity does not change with movements in the risk factor. This simplifies calculations, as we do not need to recalculate portfolio sensitivities for large moves.

Readings

RiskMetrics technical document chapters 4, 5 and 6 (150 min)

RiskMetrics 'return' chapters 1, 2 and 3 (90 min)

Hull, chapter 17 (60 min)

Key things you should know

How to calculate VaR

The random walk model

Normal distribution of returns, lognormal distribution

'Thin waist' and 'fat tails'

Measures of market risk

Properties of prices and return under the random walk model

Autocorrelation, homoscedasticity

Duration map

Cash-flow mapping

Study Questions

Q. Under the standard parametric VaR methodology, which of the following assumptions is true?

- a) Returns follow a log normal distribution
- b) Log returns follow a normal distribution
- c) Mean log return is zero for daily VaR
- d) All of the above

The values of the lognormal distribution are all positive, the normal distribution covers all values. The answer a) would mean that returns be always positive, b) would mean that we talk of logs of negative numbers. This leaves us with c)

Historical calculation of VaR, Monte Carlo

What this is about

There are different methods to compute VaR, each with their own merits. Historical calculation avoids the fat tails issue, while Monte Carlo consists of running (on a computer) a large number of scenarios based upon specified distributions. The historical approach uses historical data directly to evaluate the portfolio risk. The benefit is a realistic distribution of risk (so long as the future is like the past!). The downside is that the data set is finite and limited, so we may suffer from poor statistics. Monte-Carlo in effect generates historical price data to order with average characteristics (mean, std) matched to history or to a future scenario. Ultimately, MC is more flexible but requires a considerable amount of computer time.

Readings

Crouhy, chapter 5 (sections 2 to 5 including appendix), chapter 6 (sections 1 and 2) (40 min)

Key things you should know

Parallel shift

Historical and parametric VaR

1-day, multi-day VaR
VaR as a management tool
Monte Carlo VaR
Improvements to VaR

Study Questions

Q. When is a Monte Carlo approach most advantageous for computing VaR

- a) In volatile markets
- b) When the markets expected behaviour is non-normally distributed
- c) When quick calculations are requested
- d) When regulators are getting nervous about VaR calculations

Monte Carlo approaches provide a good flexibility in setting the distributions of returns, but require a lot of calculation time and computing power and time. VaR should be more of a management tool than a defence against regulators: b)

Covariance matrix construction

This includes UWMA (Un-Weighted Moving Average) and EWMA (Exponentially Weighted Moving Average).

What this is about

A critical factor in estimating VaR is the estimation of the parameters required by the variance-covariance matrix. Just taking past value is not enough when markets are behaving erratically. There are also mathematical “technical” factors to consider as the calculation may fail due to a slight mis-specification or data error leading to singular or non-positive definite matrices.

Readings

Crouhy chapter 6 sections 4 to 7 including appendix (30 min)
Jorion chapter 8 sections 8.2 and 8.3 (20 min)

Key things you should know

Decay factor
Variance, covariance
Daily covariance matrix
UWMA, EWMA
Back-testing

Study Questions

Q. Under the RiskMetrics cash-flow mapping method for interest rates, price volatility is required. The formula to convert yield volatility (expressed as a percentage of current yield) to price volatility is:

- a) Price Vol (σ_p) = Modified Duration (MD) \times Interest Rate (Y) \times Yield Vol (σ_y)
- b) Price Vol (σ_p) = MD \times $\frac{\sigma_y}{Y}$

- c) Price Vol (σ_p) = MD \times σ_y
- d) Price Vol (σ_p) = MD \times (1 + Y) \times σ_y

Answer b) can be discarded outright, as it makes price volatility a decreasing function of yield, a higher yield would make a bond price less volatile. Answers c) and d) would fail to scale the volatilities, would make yield volatility lower than price volatility. Answer a) remains as the only plausible one, consistent with the definition of duration as the relative change in price for a given change in yield. Please note that there is no need to know about the RiskMetrics technical document for this specific question: a)

Q. For EWMA (Exponentially Weighted Moving Average), using a decay factor of 0.94 and a tolerance level of 1% (i.e. excluding exponential weights below 1%), the effective number of data points used to estimate the covariance matrix is:

- a) 74
- b) 150
- c) 100
- d) 250

The decay factor compounds by day, hence we have: $0.94^x = 0.01$. Using natural logarithms, $x \text{Log}(0.94) = \text{Log}(0.01)$. Hence $x = 74.4$, a). We shall note that 250 appeared as the most plausible answer, as this represents one year of daily prices.

Risk limits

This part includes market risk limits, stop-loss, exposure and VaR considerations.

What this is about

On top of a sound Risk Management, institutions must develop a risk system that includes setting limits within which traders can operate.

Readings

Crouhy chapter 3 sections 4.3 and 4.4 (20 min)

Key things you should know

Risk limits
VaR-based limits

Study Questions

Q. What should happen if a successful trader consistently uses between 80% and 95% of his trading limit?

The trader has not exceeded his trading limits, so he has not transgressed, however one would expect to see his risk to rise and fall with market opportunities, and if the market had an unexpected move, he may well exceed his limits. The type of trading and risk needs to be more closely examined and the trader reminded of the need not to exceed limits, even for an unusual market move. If after review the risk / reward ratio in his trading is favourable, it may be

applicable to increase his limits so that his normal position size now represents 50-60% of limits.

Stress-testing and scenario analysis

What this is about

VaR is a risk management tool designed for normal market situations, not for stress situations. This is why it must be complemented by stress-tests and scenarios. This can use historical market events such as the crash of 1987, the Russia crisis of 1998 to stress portfolios. Scenarios are forward looking "what-if" tests on current portfolio risks, stress-tests are specific to the particular portfolio so that the moves are particularly pathological to that position.

Readings

Crouhy, chapter 6 section 3 (20 min)

Key things you should know

Stress-test
Historical replication scenario
Hypothetical one-off scenario
Worst-case scenario
Generic scenario
Yield curve shift
Limitations of VaR

Study Questions

Q. Which of the elements below argue for the use of stress-tests?

- I. Natural catastrophes
 - II. Presence of long options in the portfolio
 - III. Terrorist networks
 - IV. Fat tails
-
- a) II, III and IV
 - b) II and III
 - c) I, III and IV
 - d) All of the above

Natural catastrophes are seldom included in VaR frameworks, let alone in our thought processes. Long options immunise a portfolio against extreme movements, so they are helpful not hurtful to the portfolio. Terrorist attacks provide possibility of shocks that are potentially correlated, hence even worse against VaR. Fat tails belie normality assumptions: c).

RAROC & economic capital allocation

What this is about

An application of VaR is to allocate capital to activities that are less risky and more lucrative. Hence Risk-adjusted returns can be computed, that not only take into account returns, but also the risks incurred to achieve these returns. RAROC

stands for Risk-Adjusted Return On Capital. The most capital should be allocated to the activity with the highest RAROC. Economic capital is the amount of capital at risk at a given (usually very high) confidence interval for the activity (e.g. 99.9%) and so is directly related to a portfolio's VaR.

Readings

Jorion pp. 424-429 (15 min)

Crouhy pp. 534-547 (30 min)

Key things you should know

RORAC, RAROC

Risk-adjusted returns

Capital allocation

Economic capital

Hurdle rate

Loan-equivalent

Study Questions

Q. Which of the following is a weakness in a RAROC implementation?

- a) Recognising different types of capital (economic, regulatory)
- b) Integrating liquidity risk
- c) Informing different departments about the capital allocation strategy across departments
- d) Analysing the possibilities of RAROC-arbitrage, including tax-based arbitrages, to design it as much as possible arbitrage-free

Regulatory capital is based upon the aim to avoid systemic risk: this is not directly an internal concern (obviously the bank must however be compliant), the institution should be concerned about economic capital per se. Liquidity risk is a component of market risk, although at the end of a chain of events. The key to allocate capital should be kept as managerial information rather than used to create tensions between capital-envious departments. Raroc should be arbitrage-free: c)

Alternative risk measures

What this is about

VaR, even complemented by scenarios and stress-tests, is but a single risk tool that needs to be implemented within a sound risk and control framework of a good banking organisation. Defining roles and responsibilities is an integral part of a risk management process as is monitoring, back-testing and management oversight

Readings

Jorion chapter 21 section 21.1 (10 min)

Key things you should know

Best practices in Risk Management

Group of 30 recommendations

Study Questions

Q. Valuation of derivative positions should be done on the basis of:

- a) Accounting figure
- b) Mark-to-Market for traded instruments, cost for OTC instrument
- c) Mark-to-Model for OTC, actuarial for traded instrument
- d) Mark-to-Market or Mark-to-Model

The investment banking world breathes in MTM: Mark to Market, or to Model if there are no market prices. Purchase cost is of no relevance after the market has moved. It is critical that all positions are evaluated on the same basis if they are being risk-managed together (e.g. exchange-traded and OTC options), else the risk numbers will be meaningless. Furthermore, within a trading environment the ability to cut positions at close to market values is a vital part of the risk management process, using MTM ensures that the P&L does not widely fluctuate as the portfolio is turned over: d)

Credit Risk

This part counts for 50% of the exam module III and is 15% of the total marks for the entire PRM exam.

Types of credit risk

This includes corporate vs. sovereign risk.

What this is about

Credit risk is the risk of loss arising from a counterparty not repaying an obligation due to bankruptcy (or similar). The magnitude of credit risk is usually reflected in the size of the market yield spread of the counterparty's bonds to Government bonds.

The perception of Credit risk, the traditional risk of the banking industry, has changed rapidly in the past decade. The credit risk inherent to each debt must be assessed in a consistent way.

Readings

Jorion Chapter 3 section 3.4 and Chapter 13 sections 13.1 and 13.6 (20 min)

BIS (January 2001), part A (45 min)

BIS (July 1999). pp. 3-4 and Appendix (20 min).

Key things you should know

Credit risk
Sovereign risk
Credit exposure
Basle accord
Cooke ratio
Risk weight

Study Questions

Q. Why has the development of derivatives necessitated credit risk regulations, when these instruments are designed to deal with market risk?

Derivatives, being off-balance sheet, were poorly captured by previous regulation. However, they offer a wide potential for leverage, and involve very volatile credit risk exposures between market participants: the profit of a counterparty is the loss of the other counterparty. Derivatives, for all their merits, have blurred the distinction between credit risk and market risk. Therefore, besides their usefulness in transferring risk, they created an increase in credit risk in the system, that needed a better monitoring and assignment of risk capital.

Actuarial methods

This includes market prices, accounting (Altman) and other actuarial methods.

What this is about

Default prediction models use accounting and market data to assess the probability of default. A credit risk model must incorporate and blend accounting data, market data and ratings. These methods rely on historical default rates gathered over a long period of time so as to associate risk of default of a debtor with credit rating.

Readings

CreditMetrics Technical Document. Chapter 1 (45 min)

CreditMetrics Chapter 5 (20 min)

Key things you should know

- Portfolio-based approach
- Concentration risk
- Ratings, downgrade value
- Default state
- Distribution of value
- Probability density function
- Standard deviation, percentile level
- Expected credit loss
- Exposure type
- Z-score

Study Questions

Q. What are the main limitations of standard deviation as an indicator of credit risk?

The distribution of credit risk losses is skewed (asymmetric). Hence the standard deviation combines information related to the right-hand side of the curve (upside), which is not directly relevant to credit risk assessments. To complement standard deviation, percentiles of loss distribution can be used.

Exposure, loss given default (LGD) and expected loss

What this is about

The elements to assess for credit risk are: current and potential exposure, probability of default, recovery rate, and loss given default. Each of these steps requires analysis of accounting, market and ratings data. Exposure is usually characterised as the market cost of replacing the position, the loss given default takes into account that although default has occurred you are likely to recover some of your money from the debtor. The expected loss takes into account exposure and LGD to derive the actual expected loss. The LGD% is equal to 1-(the recovery rate%).

Readings

Jorion Chapter 13 sections 13.2 and 13.3 (20 min)

CreditMetrics Chapter 2 (20 min)

BIS (April 1999), parts I and II sections 2, 3 and 4 (40 min)

Key things you should know

Default rate
Recovery rate
Current, potential exposure
Mark-to-Market
LGD, EL, EDF
Forward zero curve
Forward value
Time horizon

Study Question

Q. How is the loss given default incorporated in the CreditMetrics Technical Document?

- a) The document does not consider it
- b) By a parameterized distribution
- c) By a look up table
- d) By a constant

Although the document does not mention the term of loss given default, this concept is handled in chapter 7 (pp. 77-80) through recovery rates. The rationale is that recovery rates are 'highly uncertain', but can be assessed through beta distributions (p. 80), with different parameters for different seniority classes: b)

Rating agencies and their grades

What this is about

Rating agencies play a key role in the banking system, in helping operators to assess obligors and their likelihoods to honour their obligations. They use methodologies based upon models and experience, which do not predict every

default on time. The highest rating is AAA and the lowest C. Different rating agents (Moody, S&P, Fitch, IBCA, etc.) use their own variations on this scale.

Readings

Crouhy Chapter 7 sections 1-4 (30 min)

Crouhy Chapter 7 sections 5-8 including appendices (30 min)

Key things you should know

Internal ratings

Measuring default probability

Ratings

Investment grade

Migration matrix

Key ratios

Financial assessment

Country risk reports

Study Questions

Q. What are the main limitations of an assessment exclusively based on accounting figures?

The Enron case highlights enough that the statement is true. In less dramatic cases, however, the following reasons can be alleged:

- accounting figures, by nature, are based on the past, while a rating is aimed at informing about future possible events
- the possibility of a firm to access cash through capital markets can change daily
- accounting figures must be used with judgement whenever possible

Settlement risk and netting systems

What this is about

Settlement risk is the risk of a payment not being received because of administrative failure. A netting system allows compensation of exposures between counterparties. As such, it considerably reduces risk, while involving numerous exposure calculations. Margined exchanges such as LIFFE go a long way to eliminate settlement risk. Herstatt Bank's failure to deliver on one leg of a FX deal is often cited as a good example of settlement risk.

Readings

Jorion Chapter 13 section 13.4 (10 min)

BIS (January 2001), parts B1, B2, B3 (45 min)

Key things you should know

Closeout netting

Effects of netting

Replacement value

Study Questions

Q. Herstatt Risk relates to:

- a) the market risk of an FX contract
- b) the German Mark debacle of 1978
- c) the settlement risk of an FX contract
- d) none of the above

Herstatt Bank, in 1974, went bankrupt following a missing payment due to different time zones in which it was operating. For this question, the candidate needs to understand what is Herstatt risk, or at least that it was related to a settlement risk.

Marginal, cumulative default risk

What this is about

Marginal, or yearly, default rates are used to calculate default probability over longer periods. Much of this has been studied under 'exposure, LGD, EL'. Cumulative default risk is the effective sum of each marginal default for each year of the period considered (e.g. 5 y cumulative default is the product of five 1 year marginal defaults).

Readings

Jorion Chapter 13 section 13.3; pp. 319-320 (10 min)

Key things you should know

Autocorrelation of default
Cumulative probability
Survival rate

Study Questions

Q. The default rates on a portfolio have been estimated at 2% for the coming year and 4% for next year. What is the expected payment of 2-year obligations?

The cumulative default rate is $(1-0.02)*(1-0.04) = 0.9408$, 94% of obligations are likely to be paid back.

Transition matrix

What this is about

A transition matrix is a practical way to represent the likelihood, over a certain period, for obligors of one rating category to migrate to another one e.g. AA to A over one year. This is an essential use of ratings for constructing credit risk models.

Readings

CreditMetrics Chapter 6 (40 min)

Key things you should know

Credit quality

Migration

One-year, multiyear matrix

Study Questions

Q. Assuming independence and a recovery rate of 70%, what is the expected loss of the following portfolio?

	Face value of the bond	Probability of default
Bond A	1,000 Euros (EUR)	0.4
Bond B	2,000 EUR	0.3

- a) 300 EUR
- b) 900 EUR
- c) 1,000 EUR
- d) None of the above

This is a straightforward calculation. The main thing is not to mix up risk exposure, loss given default and expected loss.

On bond A, we lose $1000 * (1-70\%)$ with probability 0.4, giving 120 EUR

On Bond B, we lose $2000 * (1-70\%)$ with probability 0.3, giving 180 EUR

Total is 300 EUR: a)

Q. Given a one-year probability of default of 20%, what would be the cumulative probability of default for the bond for the three years?

- a) 45.4%
- b) 48.8%
- c) 60.5%
- d) None of the above

The probability of non-default is 80%, for 3 years it becomes $0.8 * 0.8 * 0.8 = 51.2\%$, hence probability of default before 3 years is 48.8%: b)

Joint transition matrices and correlated migrations

What this is about

A downgrade, or default, sometimes comes out of the blue, sometimes due to influence of group structures. Therefore, it is useful to determine transition probability through lateral information from other companies than the obligor.

Readings

CreditMetrics Chapters 3 and 8 (to p. 93) (30 min)

BIS (April 1999), part II section 6 (10 min)

Key things you should know

Joint probability
Credit event correlation
Portfolio credit

Study Questions

Q. The portfolio contains one risky bond from company A. Company A is a subsidiary of XYZ and if XYZ defaults, company A does so too. The probability of default of XYZ is 0.3 and the probability of company A going into bankruptcy without XYZ defaulting is 0.5. What is the probability of having a default on the risky bond?

- a) Cannot be determined
- b) 0.60
- c) 0.70
- d) None of the above

It looks necessary to draw a mental map of events: if the parent defaults or not, if the subsidiary defaults or not, with their probabilities. Company A will default either following its parent, which will default with probability 30%, or if its parent does not default (probability 70%) but the subsidiary does default (probability of stand-alone default: $0.7 * 0.5 = 0.35$). Hence the probability of default of A is 65%: d).

Credit derivatives

These include default swaps and total return swaps.

What this is about

Alongside traditional methods to manage credit risk, credit derivatives, which have been dramatically developing in recent years, allow efficient and cost-effective transfer of credit risk. The Credit Default Swap (CDS) pays out if the obligor defaults and is similar to credit insurance on bonds. The total return swap allows transfer of credit risk to counterparty without the direct sale of the bond. Repackaged bonds, like Collateralised Bond Obligations (CBOs), are also frequently viewed as credit derivatives.

Readings

Crouhy Chapter 12 (60 min)

BIS (January 2001), part B4 (20 min)

Key things you should know

Credit enhancement tools
Total return swaps
Credit default swaps
Spread options
CLOs
Risk management requirements for credit derivatives
Protected exposure

Study Questions

Q. Under what common circumstances is it advisable for an AAA-rated corporate to purchase a total return swap on a client from a counterparty of a lesser rating?

- a) Any time, as credit risk is not the business of a corporate
- b) When the risk concentration on a category of clients is excessive
- c) When anticipating an upgrade for the counterparty
- d) Never, as an AAA can fund its credit risk

It generally is not advisable to transfer risk to a lesser rating, as the funding costs of the corporate are likely to be lower than those of the counterparty. There are no compelling reasons to systematically transfer or retain credit risks for a corporate, it can depend on an array of circumstances. There is little sense to speculate on banks' rating changes. It makes sense for a corporate to manage the portfolio of credit risks as a portfolio, and therefore to avoid excessive concentrations without interfering with the commercial relationship: b).

Recovery rate distributions

What this is about

Given default of an obligor, the debt is worth its face value times the recovery rate. Estimation of this rate can be predicted and approximated by a statistical distribution or historical recovery tables.

Readings

CreditMetrics Chapter 7 (15 min)

Key things you should know

Seniority classes
Distribution of recovery rates

Study Questions

Q. The distribution of recovery rates is characterised by:

- a) Fairly consistent mean, low standard deviation
- b) Fairly uncertain mean, high standard deviation
- c) Fairly consistent mean, low standard deviation
- d) Fairly uncertain mean, high standard deviation

The mean is between 0% and 100% (over long periods, 50% is about right). As for the standard deviation, uncertainty reigns even more: d)

Implied default probability

Default probabilities are inferred from corporate bond credit spreads.

What this is about

Risk-neutral valuation gives a simple relationship between the market price of an asset and the probability of default of the obligor, by arbitrage with an equivalent but risk-free instrument, and by equating expected values of cash flows. This provides the price of credit risk.

Readings

Jorion Chapter 13 section 13.5 (10 min)

Key things you should know

Expected/unexpected credit loss
Credit spread
Market price of credit risk
Credit reserve

Study Questions

Q. What is the probability of default of the issuer of a zero-coupon 1-year bond trading at 80 b.p. above the yield curve, if the expected recovery rate is 50% and risk-free interest rate is 5%?

p= probability of default
r= risk-free rate
s= spread

The expected value of the cash flows from the risky bond is: $100 * ((1-p) + p/2)$.

This discounted value, at the rate of $1+r+s$, equates that of the risk-free bond.

$100*(1+p/2)/(1+r+s)=100/(1+r)$. Hence $p=1.5\%$

Merton and KMV models

What this is about

Different institutions developed models to deal with portfolio credit risk, in modelling correlation. CreditMetrics, CreditRisk+ and CreditPortfolioView (respectively J.P. Morgan, Credit Suisse and Mc Kinsey) offer different approaches towards credit VaR calculations.

Readings

Crouhy Chapter 9 Section 5 pp. 368-389 "The KMV Approach" (45 min)

Jorion Chapter 13 section 13.7 (10 min)

BIS (April 1999), part II section 5 (10 min)

Key things you should know

The main credit risk models, their assumptions and different approaches
The KMV approach
Distance to default
Risk-neutral EDF

Study Questions

Q. What is characteristic of a default mode credit risk model?

- a) A model which considers two states of nature
- b) A model which incorporates a default definition
- c) A model which considers default as a reflecting state
- d) None of the above

It helps here to understand default mode as the opposite of ratings-based approach. Default mode models consider only default and non-default: a)

Q. What is the one-year transition matrix assuming only two categories [*i.e.* default (d) and non-default (nd)] from a portfolio with 300 loans and the following payment history?

Number of loans that transited from non default to default in 2000										
Jan	Feb	March	Apr	Jun	Jul	Aug	Sep	Oct	Nov	Dec
0	9	0	0	6	0	3	3	0	6	3

a)

.7	.3
0	1

b)

.9	.1
0	1

c)

.9	.1
.1	.9

d)

.3	.7
1	0

From the table above, 30 (10% borrowers) defaulted. The transition matrix indicates the probability for an exposure to go from non-default (0) to default (1): this discards a) and d). The c) matrix does not indicate the statuses 0 and 1, and makes little sense. The b) matrix does: b)

The Merton (1974) model implies that a position in a credit-sensitive bond is equivalent to:

- a) A long position in the firm's equity and a short position in a risk-free bond
- b) A long put and a long call position on the firm's assets
- c) A long position in a credit-risk-free bond and a short put on the firm's assets
- d) An up-and-in call on a credit-risk-free bond and a short call on the firm's equity

A long position on a bond means a short cash position. If the bond issuer defaults, the bondholder is left with a bad loss, hence the position equates to shorting something. Hence only c) remains. A more thorough way to address this question is as follows:

- a short position on a risk-free bond means a borrowing; this is not the case of a long bond position: a) falls
- such a position as in b) would make the put very valuable in case of issuers' default: b) falls
- if the firm's (net) assets are worth nothing, a short put represents a loss (as with a risky bond)
- an up-and-in call (besides the fact that this exotic was unknown in 1974) gives the holder, after the bond has gone higher than a certain level, the right to purchase the bond: this does not equate a risky bond.

RAROC & economic capital allocation

What this is about

Regulatory capital and economic capital are calculated on the basis of the degree of risk of an activity, for external or internal purposes. The following texts show an overview of the issues linked to using credit risk tools for an optimal capital allocation.

Readings

Jorion Chapter 13 section 13.8 (10 min)

BIS (January 2001), parts B5, B6, B7 and annexes (30 min)

BIS (April 1999), part II section 1 (10 min)

Key things you should know

Integration of credit and market risk

Credit risk of derivatives

Haircut

Mismatches

Study Questions

Q. Which of the following is an appropriate way to measure operational risk?

- a) VaR
- b) Notional exposure
- c) Loss data distribution
- d) Insurance values

VaR is a market risk management tool, notional exposures are related to credit risk, insurance values show only insured assets instead of operations. Loss distribution data record the history of operational losses: c),

Operational Risk

This part counts for 17% of the exam III and is therefore 5% of the total marks for the whole PRM exam.

Operational risk is a discipline in the process of codification. On some of the topics below, there is very little in terms of commonly accepted standards. In every institution, different methodologies are used. The Basle Committee of Banking Supervision provides the closest thing so far to a standard. Read the text from Basle Committee, Consultative document on Operational Risk, p. 1-14 (60 min). This text, aimed at preparing the design of a risk-based regulation rather than directly managing operational risk, refers to the following sections indirectly.

Typologies of operational risk

What this is about

Operational risk is an ill-defined, uneasy to capture type of risk. Several definitions and typologies are used, most of which have their merits.

Readings

Jorion Chapter 19 sections 1 to 3 (10 min)
Crouhy Chapter 13 sections 1 and 2 (15 min)

Key things you should know

Definitions of operational risk
Bayesian approaches vs. resilience approach
Identification
Assessment and measurement
Prevention and mitigation
Capital attribution

Study Questions

Q. Which one of the following is a risk driver rather than a risk indicator?

- a) Staff turnover
- b) Product complexity
- c) Systems downtime
- d) Model errors

Indicators are response variables, while drivers are decision variables, so the answer is b)

Insurance, reinsurance

What this is about

Insurance and banking risks are converging types of risk. The Basle Committee recognised insurance as a risk mitigation factor. Reinsurance works by slicing up the liability under an insurance contract (e.g. first USD 10m loss, next USD 100m loss etc.) and selling it on to other companies for a proportion of the premium. The loss therefore is both syndicated (split between underwriters) and reinsured by tranching off the very high loss very low probability (tail) risk to specialist re-insurers.

Readings

Basle Committee. Operational Risk. p. 15. (10 min)

Key things you should know

Risk mitigation
The role of insurance
Re-insurance

Study Questions

Q. What is the main hurdle seen by the Basle Committee in fully recognising the use of insurance?

The market for insurance products for banking operations is 'still developing', which means the products are not yet widespread enough and not standardised. Besides, quantification is arduous. Another reason is that the presence of insurance may replace an operational risk with a counterparty risk. Insurance companies do not usually pay up "on the nail" , they generally seek to reduce a claim through loss adjustment and litigation and hence have a different "risk model" from the banking sector.

Causal models

What this is about

An approach to assess operational risk is to try to predict what can happen and the consequences in terms of losses, in reconstituting the chain of events. In essence "what can go wrong, how will it effect us?"

Readings

Crouhy Chapter 13 sections 3 and 4 (15 min)

Key things you should know

Bayesian approach
Cause and effect
Operational process
Resilience
Likelihood of occurrence
Potential impact

Study Questions

Q. Which of the following principles does not help in an operational risk measurement process:

- a) Consistency
- b) Transparency
- c) Timeliness
- d) Relevance

Risk exposures should be adequately reported to senior management. Consistency and relevance make reporting easier to use, transparency ensures that information is truthfully reported. Timeliness is less relevant (contrarily to market risk), as an existing risk can be present for a long time without occurring.

Risk management processes

These include prevention, mitigation and insurance.

What this is about

Operational risk can be tackled at different stages of their occurrence and with different methods. An explicit process must be put in place for managing operational risk.

Readings

Jorion Chapters 18 (30 min)

Crouhy Chapter 13 section 5 (20 min)

Key things you should know

Risk assessment

Risk reporting

Risk categories

Sources of risk

Study Questions

Q. When used to protect against catastrophic risks, Insurance:

- a) Reduces the need for capital by more than 50%
- b) Transforms catastrophic risk into counterparty risk
- c) Is always too expensive, as actuaries price to a certain return for the insurance companies
- d) Eliminates default risk

Insurance is a way to transform a more esoteric and less measurable risk into something more widely understood like counterparty risk, c).

Loss events databases and their uses

What this is about

Quantitative models need data to feed into. This is why collecting loss data is a critical exercise for the calculation of Operational Risk measures and for capital allocation

Readings

Jorion Chapter 19 section 4 (15 min)

Key things you should know

Internal/external data

High/ low frequency events

High/ low impact events

Study Questions

Q. What are the main advantages of using external loss databases?

- a) Access to a wider pool of data
- b) Potential access to competitors' data
- c) Access to well-structured data

d) Access to regulation-compliant data

The data in external databases are edited so that names and other means of identification of the origin are deleted. Moreover, these databases are not designed for mutual spying, but for common progress. Every institution should design and structure these databases at the outset so that they can stand the test of time as well as external databases. A badly structured internal database is likely to be a costly and useless exercise. Internal databases should be, at the outset, regulation-compliant, as regulation is flexible enough to allow tools that are compliant as well as internally useful. For high-impact low-frequency data, internal data are likely to be too succinct. The collated data of several institutions are likely to be a much better guide to the future: a)

RAROC & economic capital allocation

What this is about

Operational risk can be avoided, monitored, prevented or mitigated, but a residual risk will strike sooner or later. This is why operational risk must be tackled at every stage in the cause-effect relationships, including with capital reserves for operational losses.

Readings

Jorion Chapter 19 sections 5 and 6, and chapter 20 (30 min)

Crouhy Chapter 13 sections 6, 7, 8 and 9 including appendices (20 min)

Key things you should know

Expected/unexpected losses

Risk retention and transfer

Business risks

Events risks

Reputational risks

Integrated risks

Natural hedge

Study Questions

Q. What could be the most effective hedge of a portfolio of weather derivatives?

- a) Back-to-back matching
- b) Global diversification
- c) Catastrophe bonds
- d) Equity sector equity index futures

Weather derivatives are not likely to find perfect hedges, as these instruments are new, fragmented and thinly traded. Back-to-back matching (buying protection from winter sports resorts, selling to beach resorts, or similar approaches) is of limited application. Hedging with a position on a particular sector, if there were enough of these instruments, would be similarly insufficient. Catastrophe bonds would matter only for extreme weather. Global diversification can help compensate the effect of dry weather in a region with excessive rain in another one: b)

Exam IV

This exam contains: Case Studies and PRMIA Standards of Best Practice, Conduct and Ethics and PRMIA's Bylaws. It lasts 1 hour, and counts for 25% of the total PRM exam points. The total time estimated to read the source texts for this module is 7 hours.

In spite of the Risk Management techniques we have seen in the previous exams, disasters have happened, and will happen again. Finance is a risky business by nature. Therefore, it is critical to recognise a disaster waiting to happen when we see one. For this, the past may be a guide for the future. Risk management also requires an emphasis on adherence to high ethical standards. Finally, all risk managers who choose to "associate" need to understand their own corporate governance structure.

Case Studies

This part counts for 80% of the exam module IV and hence represents 20% of the total marks for the whole PRM exam.

Barings

What this is about

In 1995, the UK's Barings Bank was brought to the point of collapse and was absorbed by ING for £1, following USD 1.3 BN losses on futures from the Singapore office. Nick Leeson, the responsible trader, was tried and convicted of fraud. Many of the issues could have been avoided by proper internal controls and segregation of duties and responsibilities. This accelerated the global trend of the financial industry towards more risk management.

Readings

Barings

Jorion Chapters 2 (Section 2.2.1) and 21 (Section 21.1.2) (20 min)

Bank of England. Report on the Collapse of Barings Bank, 1995 (45 min)

International Financial Risk Institute. Not Just One Man-Barings. (20 min)

Key things you should know

The organisation of Barings' Singapore office

The chain of events that broke up Barings

The role of Nick Leeson

The lessons from the events

Study Question

Q. Which position would have partially hedged Nick Leeson's primary option position at Barings?

- a) Long futures
- b) Short Strangle
- c) Long Strangle
- d) Total Return Swap

The option positions were due to sales of puts and calls on the Nikkei 225 index. A long strangle is a position that gives a positive return if the price of the underlying goes up or down, a small loss if not (V shape, but with a flat bottom). Total return swaps, which are credit derivatives, apart from the fact that they did not exist at this time, would not have protected him. Long futures would not have protected in case of drop in the index: c)

Q. Nick Leeson tried to hide his losses using what method?

- a) Portaling
- b) Switching
- c) Re-margining
- d) Volatility Smiles

Nick Leeson hid his positions into an unreported account 88888, so as not to report them to head office : b)

Metallgesellschaft

What this is about

In 1993, the German industrial conglomerate Metallgesellschaft had, through its subsidiary MGRM, losses of USD 1.3 BN due to a hedge on oil positions that showed unrealised losses during the lifetime of the hedge. The parent company closed the positions. An interesting but controversial debate has been going on about the responsibilities in the losses, as some contend that the hedge would have brought a profit in the end, while some think these positions were not sustainable. Essentially, part of the problem was liquidity and position size. Losses on futures hedges, via the margin call, were realised everyday whereas the large unrealised gain on OTC swaps gave no offsetting daily cash flow. The position and funding requirements became "too big to hold" and forced a fire sale at a loss.

Readings

Jorion Chapter 2 (Section 2.2.2) (15 min)

Digenan, Felson, Kelly and Wienert. Metallgesellschaft AG: A Case Study. (20 min)

Krapels, 'Re-examining the Metallgesellschaft Affair and its Implication for Oil Traders'. (30 min)

Key things you should know

Governance issues at Metallgesellschaft
Hedging policies and strategy
The risks involved in the hedge

Study Questions

Q. What caused the losses for Metallgesellschaft?

- a) At the final maturity date the price in the futures was well below the market price

- b) To hold the position, they assumed a constant interest rate to invest the proceeds
- c) At the final maturity date the price in the futures was well above the market price
- d) To hold the position, they assumed an unbounded pool of resources

What went wrong was during the hedge. Answers a) and b) assume this was not a hedge but a speculation. Interest rates played a lesser role in the story, while the covering of the unrealised losses during the operation was what made the parent company to unwind the positions, judged to show too deep losses: d)

In the Metallgesellschaft case, what was the communication problem between the subsidiary and the parent company?

- a) They did not communicate the loss as soon as it started to kick in
- b) They did not communicate the intrinsic bet on the price of gas
- c) They did not explain the economics of the strategy
- d) They did not explain why they had to buy more future contracts

The hedge, over 5 to 10 years, were a non-trivial piece of financial engineering. The parent company was more scared by the level of losses showed by the position than ready to listen to the rationale of the operation: c)

Long Term Capital Management

What this is about

In 1998, the hedge fund LTCM made losses on highly leveraged positions. No new investors popped in. When funding was not available and positions proved difficult to liquidate, the New York Fed encouraged a pool of banks to invest money into the venture, to protect the financial system. The event was all the more unexpected because LTCM was managed by a very high-profile team. The major positions were in bond swap spreads, high leveraged through the use of prime brokerage accounts for bond repo at very low (or zero) haircut rates.

Readings

Jorion Chapters 14 (Section 14.4) and Chapter 21 (Sections 21.1.3 and 21.4) (30 min)

United States Treasury. Hedge Funds, Leverage and Lessons from Long Term Capital Management, pp. 1-42 (60 min)

Shirreff. Lessons from the Collapse Of Hedge Fund, Long-Term Capital Management. (20 min)

Key things you should know

Hedge funds setup
When genius fails
Funding liquidity risk
Integrated view of risk

Study Questions

Q. The strategy of getting creditors to lend money and invest equity in LTCM had the effect of:

- a) Increasing transparency to the creditors
- b) Reducing leverage
- c) Giving LTCM partners a put on the value of the fund
- d) No relevant effect as equity and debt offset each other

Transparency is driven by information flows, not cash flows. The investment in equity came to offset the losses, leverage was not reduced as compared to the initial (pre-loss) situation. The intervention of investors did not increase the positions, and added capital, to an effect. A bank took up shares and sold LTCM a call option, resulting in the investors effectively holding a put on LTCM shares: c)

Q. LTCM's balance sheet as of August 31, 1998 showed the following (USD):

- a) 100 billion in assets, -0.5 billion in equity
- b) 125 billion in assets, 2.3 billion in equity
- c) 400 billion in assets, 4.0 billion in equity
- d) 125 billion in assets, 6.1 billion in equity

Although answer a) would probably have meant liquidation, this is really a question of knowing the degree to which LTCM leveraged their balance sheet, which was approximately 60:1: b).

Group of 30 Report

What this is about

In 1993, a consultative group of bankers issued a set of rules and recommendations for using derivatives, setting the scene for regulation. These outlined the principles without that protect financial institutions from various risks.

Readings

Jorion Chapter 2 (Section 2.3.1) and chapter 21 (section 21.1.1) (10 min)

Group of Thirty, Derivatives Practices and Principles (20 min)

Key things you should know

Set of management practices
The contribution of derivatives to risk

Study Questions

Q. According to the G-30, derivative credit exposure should be measured by:

- a) Current Exposure
- b) Potential Exposure
- c) a) plus b)
- d) a) plus b) minus Posted Collateral

This is principle 10, although collateral is mentioned in recommendations 7 and 14. Current exposure includes the credit exposure should the position be

unwound, potential exposure includes exposure due to future market fluctuations. Collateral protects its owner against the impact of credit risk: d).

Q. According to the G-30 report, an ISDA master agreement is:

- a) Sufficient to prevent loss from counterparty default
- b) Not substantially enhanced by a netting provision as bankruptcy courts widely recognize netting as a best practice
- c) Enhanced when multiple master agreements exist between the same counterparties so that the legal risk of an oversight in documentation is reduced
- d) None of these

ISDA produces master agreements, which are standard documentation for derivatives. These help market participants to address contract and documentation risks. ISDA and G30 are not directly connected. Answer a) would be too good to be true (fancy not having to worry about counterparty default?). A netting provision must be explicit, courts in most countries would not take these for granted. Answer c) would defeat the purpose of master agreements, which is to standardise contracts as much as possible. Besides, G30 would not comment specifically about ISDA's work, but would rather comment on the use of standard documentation: d)

PRMIA Bylaws and Code of Conduct

This part counts for 20% of the exam module IV and so is 5% of the total PRM exam.

What this is about

The development of the Risk Management profession needs an organisation that promotes sound practices, raises levels of competence and ethics. PRMIA, a tax-exempt, non-profit organisation founded in 2002, aims at promoting sound Risk Management practices. It is the duty of Risk Managers to know and abide by the ethical rules and code of conduct set by PRMIA.

Readings

PRMIA Standards of Best Practice, Conduct and Ethics (50 min)

PRMIA Bylaws (40min)

These texts, especially the first one, require a thorough reading, for the member of the Risk Management profession to be fully aware of the duties inherent to the profession and, to understand the governance of their own association

Key things you should know

Professional behaviour expected from members

Ethical behaviour expected from members

Membership of PRMIA

Functioning of PRMIA

The development of the professional Risk Management profession through its association

Study Questions

Q. Which of the following are excluded from being Regular Members of PRMIA by PRMIA Bylaws?

- a) Corporations
- b) Corporations and former Risk Managers
- c) Regulators
- d) Students

Q. Regular members (art 3.1.1) are individual persons. Corporations, associations and regulators can become affiliate members. Former Risk Managers, as well as current Risk Managers and students, are welcome: a).

Which of the following is NOT part of PRMIA's guidance on Best Practices?

- a) Only standard methods of assessing risk should be used
- b) PRMIA members must possess, be under the supervision of someone who possesses, or inform their supervisor of the lack of required skills and/or certification to complete their risk assessment work.
- c) PRMIA members must not intentionally deceive others
- d) PRMIA members must value validation of their work by peers

Innovation is always required in the new discipline that is Risk Management. Competence and ethics are always required in financial professions. The Risk Management profession progresses because owing to discussion with peers and cross-validation of work. Fortunately, PRMIA does not prescribe orthodoxy, and encourages members to use sound practices rather than use, let alone blindly use, standard methods: a)

Appendix: further reading

The present list can be used by candidates who want to study a subject more in depth. Some of the texts below are available by download from the web addresses as indicated, some can be found in most business and university libraries.

Anton, Howard et al. (2001). *Calculus*, Seventh Edition, New York: Wiley & Sons.

Bank of England (1995). *Report of the Board of Banking Supervision Inquiry into the Circumstances of the Collapse of Barings*, London: HMSO.
<http://www.numa.com/ref/barings/bar00.htm>

Basle Committee on Banking Supervision (2001). *The Standardised Approach to Credit Risk* <http://www.bis.org/publ/bcbsca04.pdf>

Basle Committee on Banking Supervision (2001). *The Internal Ratings-based Approach* <http://www.bis.org/publ/bcbsca05.pdf>

Basle Committee (2003). *Sound Practices for the Management and Supervision of Operational Risk*
<http://www.bis.org/publ/bcbs96.pdf>

Best, Phillip (1998). *Implementing Value at Risk*, New York: Wiley & Sons.

Black, Fischer, and Scholes, Myron (1973). *The Pricing of Options and Corporate Liabilities*, *Journal of Political Economy*, Vol. 81, pp. 229-246. This is an essential text for the basic principles of options.

Abstract: If options are correctly priced in the market, it should not be possible to make sure profits by creating portfolios of long and short positions in options and their underlying stocks. Using this principle, a theoretical valuation formula for options is derived. Since almost all corporate liabilities can be viewed as combination of options, the formula and the analysis that led to it are also applicable to corporate liabilities such as common stock, corporate bonds, and warrants. In particular, the formula can be used to derive the discount that should be applied to a corporate bond because of the possibility of default.

Bollen, Nicholas P. and Robert E. Whaley (1998). *Simulating Supply*, *Risk*, 11(9), pp. 143-17.

Breeden, D. T. (1979) *An Intertemporal Asset Pricing Model With Stochastic Consumption and Investment Opportunities*, *Journal of Financial Economics*, 7:265:96.

Abstract: This paper derives a single-beta asset pricing model in a multi-good, continuous-time model with uncertain consumption-goods prices and uncertain investment opportunities. When no riskless asset exists, a zero-beta pricing model is derived. Asset bets are measured relative to changes in the aggregate real consumption rate, rather than relative to the market. In a single-good model, an individual's asset portfolio results in an optimal consumption rate that has the maximum possible correlation with changes in aggregate consumption. If the capital markets are unconstrained Pareto-optimal, then change in all individuals' optimal consumption rates are shown to be perfectly correlated.

Briys, Eric, et. al. (1998). *Options, Futures and Exotic Derivatives: Theory, Application and Practice*, New York: Wiley & Sons.

- Butler, Cormac (1999). *Mastering Value at Risk: A Step-by-Step Guide to Understanding and Applying VaR*, London: Financial Times.
- Caouette, John B, et al. (1998). *Managing Credit Risk: The Next Great Financial Challenge*, New York: Wiley & Sons.
- Chriss, Neil A. (1997). *Black-Scholes and Beyond: Option Pricing Models*, New York: McGraw-Hill.
- CreditMetrics Technical Document (1997): <http://riskmetrics.com>
- Cruz, Marcelo G. (2002). *Modelling, Measuring and Hedging Operational Risk*, New York: Wiley & Sons.
- Culp, Christopher L. (2001). *The Risk Management Process: Business Strategy and Tactics*, New York: Wiley & Sons.
- Culp, Christopher L., and Merton H. Miller (1995). *Metallgesellschaft and the Economics of Synthetic Storage*, *Journal of Applied Corporate Finance*, 7 (4), pp. 62-76.
- Culp, Christopher L., and Merton H. Miller (1999). *Corporate Hedging in Theory and Practice: Lessons from Metallgesellschaft*, London: Risk Books.
- DeGroot, Morris H. (1986). *Probability and Statistics*, Second Edition. Reading: Addison Wesley.
- Douglas, Livingston G. (1990). *A Guide to Duration and Convexity*. New York: New York Institute of Finance.
- Dowd, Kevin (1998). *Beyond Value at Risk*, New York: Wiley & Sons.
- Dunbar, Nicholas (1999). *Inventing Money: the Story of Long-Term Capital Management and the Legend Behind It*, New York: Wiley & Sons.
- Edwards, Franklin R., and Michael S. Cantor (1995). *The collapse of Metallgesellschaft: Unhedgeable Risks, Poor Hedging Strategy, or Just Bad Luck?*, *Journal of Applied Corporate Finance* 8(1), pp. 86-105.
- Elton, Edwin J., and Martin J. Gruber (2002). *Modern Portfolio Theory and Investment Analysis*, Sixth Edition, New York: Wiley & Sons.
- Evans, Merran, Nicholas Hastings and Brian Peacock (2000). *Statistical Distributions*, Third Edition, New York: Wiley & Sons.
- Fabozzi, Frank J. (1997). *Fixed-Income Mathematics*, Third Edition, Chicago: Irwin.
- Fay, Stephen (1997). *The Collapse of Barings*, New York: Norton.
- Federal Reserve (2002). *Trading and Capital-Markets Activities Manual*, <http://www.federalreserve.gov/boarddocs/supmanual/trading/trading.pdf>
- Federal Reserve Bank of Richmond (1998). *Instruments of The Money Market*, <http://www.rich.frb.org/pubs/instruments/>

- Gleason, James T. (2000). *Risk: The New Management Imperative in Finance*, New York: Bloomberg Press.
- Hilliard, Jimmy E. (1999). Analytics underlying the Metallgesellschaft hedge: Short-term futures in a multiperiod environment, *Review of Quantitative Finance and Accounting*, 12(3), pp. 195-219.
- Johnson, Dallas E. (1998). *Applied Multivariate Methods for Data Analysts*, Pacific Grove: Duxbury Press.
- Jorion, Philippe (2000) Risk Management Lessons from Long-Term Capital Management, <http://www.gsm.uci.edu/~jorion/papers/lbcm.pdf>
- Kern, Markus, and Bernd Rudolph (2001). Comparative Analysis of Alternative Credit Risk Models, http://www.ifk-cfs.de/papers/01_03.pdf
- King, Jack L. (2001). *Operational Risk: Measurement and Modelling*, New York: Wiley & Sons.
- Krapels, Ed (2001). Re-examining the Metallgesellschaft affair and its implication for oil traders, *Oil and Gas Journal*, <http://www.esai.com/pdf/Re-Examining%20the%20Metallgesellschaft%20Affair.pdf>
- Leeson, Nick (1996). *Rogue Trader*, London: Little Brown.
- Levich, Richard M. (2001). *Investments*, Second Edition, New York: McGraw-Hill.
- Lowenstein, Roger (2000). *When Genius Failed: The Rise and Fall of Long-Term Capital Management*, New York: Random House.
- Mello, Antonio S., and Parsons, John E. (1995). Maturity structure of a hedge matters: Lessons from the Metallgesellschaft debacle, *Journal of Applied Corporate Finance*, 8(1), pp. 106-120.
- Merton, Robert C., (1973A), "Theory of Rational Option Pricing," *Bell Journal of Economics and Management Science*, Vol. 4 (Spring), pp. 141-183.
- Merton, Robert C., (1973B), "An Intertemporal Capital Asset Pricing Model," *Econometrica*, vol. 41 (September), pp. 867-887, Reprinted in *Continuous Time Finance*, 1990, Cambridge, MA, Basil Blackwell as Chapter 15.
- Abstract: An intertemporal model for the capital market is deduced from the portfolio selection behaviour by an arbitrary number of investors who act as to maximize the expected utility of lifetime consumption and who can trade continuously in time. Explicit demand functions for assets are derived, and it is shown that, unlike the one-period model, current demands are affected by the possibility of uncertain changes in future investment opportunities. After aggregating demands and requiring market clearing, the equilibrium relationships among expected returns are derived, and contrary to the classical capital asset pricing model, expected returns on risky assets may differ from the riskless rate even when they have no systematic or market risk.
- Natenberg, Sheldon, (1994), *Option Volatility and Pricing: Advanced Strategies and Techniques*, Chicago: Probus
- Ong, Michael (1999). *Internal Credit Risk Models*, London: Risk Books.

- Pirrong, Stephen Craig (1997). Metallgesellschaft: A prudent hedger ruined, or a wildcatter on NYMEX?, *Journal of Futures Markets*, 17(5), pp. 543-578.
- Questa, Giorgio S. (1999). *Fixed-Income Analysis for the Global Financial Market: Money Market, Foreign Exchange, Securities, and Derivatives*, New York: Wiley & Sons.
- Ramos, Jose A., Soler, et al. (2000). *Financial Risk Management: A Practical Approach for Emerging Markets*, Baltimore: John Hopkins University Press.
- Rawnsley, Judith H. (1995). *Total Risk*, New York: Harper Business.
- Reilly, Frank K., and Brown, Keith C. (2000). *Investment Analysis and Portfolio Management*, Sixth Edition, Orlando, Florida: The Dryden Press.
- Ross, S. A. (1976), "Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory*, 13, December, pp. 343-362.
- Ross, Sheldon (1995). *A First Course in Probability*, Fifth Edition. Prentice Hall.
- Saunders, Anthony, and Allen, Linda (2002). *Credit Risk Measurement: New Approaches to Value at Risk and Other Paradigms*, 2nd Edition, New York: Wiley & Sons.
- Sharpe, William F., Alexander, Gordon J. and Bailey, Jeffery V. (1999). *Investments*, Sixth Edition, New Jersey: Prentice Hall.
- Shirreff, David (Undated). <http://risk.ifci.ch/146480.htm>
- Smithson, Charles W., and Smith, Clifford W. (1998). *Managing Financial Risk*, Third Edition, New York: McGraw-Hill.
- Stewart, James (1999). *Calculus*, Fourth Edition, Pacific Grove: Brooks/Cole.
- Stigum, Marcia (1991). *The Money Market*, Third Edition, New York: McGraw-Hill.
- Strang, Gilbert (1988). *Linear Algebra and Its Applications*, Third Edition, Fort Worth: Harcourt Brace.
- Sundaresan, Suresh M. (2002). *Fixed Income Markets and their Derivatives*, Second Edition, Cincinnati: South-Western.
- Sydsaeter, Knut, and Hammond, Peter J. (1995). *Mathematics for Economic Analysis*, Prentice Hall.
- Thomas, George B, Jr. and Finney, Ross L. (1996). *Calculus and Analytic Geometry*, Ninth Edition, Reading: Addison-Wesley.
- US Treasury (1999). *Hedge Funds, Leverage, and the Lessons of LTCM*.
<http://www.ustreas.gov/press/releases/docs/hedgfund.pdf>
<http://www.federalreserve.gov/boarddocs/testimony/1999/19990506.htm>
- Walmsley, Julian (2000). *The Foreign Exchange and Money Markets Guide*, Second Edition, New York: Wiley & Sons.

PRM Candidate Guidebook

2004



Updated September 1, 2004

© 2004 by the Professional Risk Managers' International Association.
All Rights Reserved.

Contents

Executive Summary	1
Program Design	2
Program Dates and Locations.....	4
Fees & Registration Details	5
Taking the Exams	6
PRM Syllabus.....	8
How to Prepare for the Exams.....	11
FAQ	13
Sample Exam Questions.....	17

Executive Summary

The PRM program is a series of evaluation exams and self-study materials, designed for the development of professional risk managers.

The PRM has been built by a broad coalition of industry leaders, all PRMIA members, to reflect the mission objectives of the association and to establish the leading form of education, validation and certification in the Risk Management profession.

The **learning objectives** for the PRM designation are to demonstrate knowledge and understanding of:

- the classic finance theory underpinning risk management
- the mathematical foundations of risk measurement
- the foundation of option theory
- financial instruments and their associated risks
- the daily form and function of trading markets
- risk management practices
- failed systems and practices from major risk events
- the PRMIA Standards of Best Practice, Conduct and Ethics
- PRMIA's Bylaws that ensure member control of the association

PRM Exams

To meet with the needs of busy professionals, the program provides for great flexibility, offering one six-hour exam for those wishing to complete evaluation in one day, or four separate exams, varying in length from one to two hours, which can be taken in any order over a period of up to two years. Passing all four exams leads to the award of the PRM designation.

Exam	Exam Name	No. of Questions	Time Allowed
I	Finance Theory, Financial Instruments and Markets	30	1.5 hours
II	Mathematical Foundations of Risk Measurement	24	2 hours
III	Risk Management Practices	36	1.5 hours
IV	Case Studies; PRMIA Standards of Best Practice, Conduct & Ethics; PRMIA Governance	30	1 hour

Cross-Over Requirements for other Designations

The program also recognizes the achievements of those who have other professional designations and gives partial credit towards completion of the requirements for the PRM designation. The "cross-over" exam requirements are:

CFA Cross-over	Exams III and IV
CIIA Cross-over	Exams III and IV
Actuarial Fellow Cross-over	Exams III and IV
CQF Cross-over	Exams III and IV
Actuarial Associate Cross-over	Exams I, III and IV
CSI Financial Risk Mgr Cross-over	Exams II, III and IV
CAIA Cross-over	Exams II, III and IV

Program Design

The Professional Risk Manager program has been designed for those:

- seeking professional certification
- looking to develop their skills or those of their staff
- looking for skills assessment of current employees
- looking for skills assessment of potential employees

The subject matter for the PRM program has been chosen based on relevance as well as the ready availability of literature on the specific subjects.

Because we are a diverse association, with [members](#) from more than 105 countries, from a variety of disciplines and varying areas of interest, the members of the [Education and Standards Committee](#) along with our [Academic Advisory Council](#) have taken the job of creating the framework to help us meet both the demands of current members as well as developing the flexibility to grow and change together with our industry. Specifically, the program has been designed with four of our [Mission](#) objectives in mind:

To be a leader of industry opinion and a proponent for the risk management profession

By setting the gold standard of development and certification for the risk management profession, we create standards of practice and accountability that will define what Professional Risk Managers represent to their colleagues and employers.

To drive the integration of practice and theory

Neither theory nor practice will successfully develop to their potential without a strong interaction between professionals in each area. The Professional Risk Manager program contains subject matter drawn from risk management theory, finance theory, the math underpinning risk measurement and applied question matter from the practices of our profession. We will directly test for knowledge of Finance (Portfolio Theory, Asset Pricing Theory, Option Theory) and basic Math (Calculus, Linear Algebra and Probability). Such knowledge is essential for risk managers as it forms the foundation of risk measurement and management.

We will also test for practical knowledge about markets: trading practices, intermediaries, settlement and other conventions in specific markets. These market choices will reflect the most global of markets as well as some more narrowly defined markets, with the latter giving us the opportunity to expose people to markets in which they may not normally be active.

Finally, to test for broad understanding of important risk management lessons, we have included a section based on the time-tested case-study approach to learning. Narratives of historical incidents where critical risk management issues have arisen give us the opportunity to conduct a sort of pathology of events from around the globe and to test members' understanding of their root causes.

To be global in our focus, promoting cross-cultural ethical standards, serving emerging as well as more developed markets

The content of the PRM program is designed to be geographically neutral. Specific sections focusing on "markets" are designed so that a variety of markets and market lessons will be studied.

The PRMIA Standards of Best Practice, Conduct and Ethics (Code of Conduct) represents the standards of behavior that Professional Risk Managers promise to their colleagues and employers. It has been written to cut across geographic and cultural boundaries. All Professional Risk Managers will be required to study the Code for the program and adhere to it to stay in good standing as a certified PRM.

Transparent, nonprofit, independent, member-focused and member-driven

PRMIA is a tax-exempt, non-profit professional association under the full control of its members. It does not have for-profit subsidiaries that may financially benefit individuals. Its [bylaws](#) provide strong protection for member rights and the assets of the membership. Net revenues from the Professional Risk Manager program benefit all members of PRMIA by being dedicated to support local chapter activities and the delivery of web-based resources. We are "of the risk professional, by the risk professional and for the risk professional."

The program has been designed by industry leaders, all PRMIA members. Any PRMIA member can submit questions for possible inclusion on the exams. Those submissions go through a quality control screening by a number of members of the Education and Standards Committee and, if approved, will make it to our exam question database. By emphasizing peer input and peer review, this approach sets a standard of measurement created by the industry.

Candidates will study the Bylaws of PRMIA, so that they are familiar with our structure and codification of member-leadership.

Be sure to read this entire document to view important information that will help you to be a successful candidate. We hope that you will join us in promoting the PRM program as the standard for our industry and that you will join in our efforts to ensure that it will continue to meet the needs of our members and our industry.

[\(Back to the Contents\)](#)

Program Dates and Locations

For your convenience, the exams are offered on every business day of the week.

You may request to schedule your exams at any time. There is no advantage to taking your exams on any particular date. Questions within exams are drawn from a large database of questions and are administered randomly, creating thousands of unique exam forms, all of comparable difficulty.

PRMIA uses the services of VUE, part of the Pearson Publishing family, to administer the PRM exams. Pearson VUE is a professional testing firm that has nearly 4,000 testing centers in more than 140 countries around the world.

Please note some important details about registration, cancellation, identification requirements and space availability in the Registration section below.

[\(Back to the Contents\)](#)

Fees & Registration Details

Fees

The program allows for you to take one, two, three or all four exams at the same time. There are discounts for taking more than one exam at a time. Payment by credit card is required.

2004

One Exam	US\$150
Two Exams	US\$235
Three Exams	US\$295
Full Program	US\$345

Registration

Registration for the PRM is handled using our [online form](#) or via fax.

To register, please follow the following steps:

- Step 1** Locate your PRMIA username. You will need this to register.
 - Step 2** Locate your first and second preferences of testing locations at <http://www.pearsonvue.com/prmia/locate>
 - Step 3** Identify a first, second and third choice of testing dates
 - Step 4** Register online at <http://www.prmia.org/register.php>
- You will need to have a credit card available to reserve your place. (Contact support@prmia.org if you require alternative payment arrangements)
 - You will receive a confirmation e-mail once the appointment has been scheduled. It is important to retain this notice.

Please note that you may only take the same exam once every ninety days. If you repeat the exam, your second score will not count and you will be charged the full fee for both exams.

[\(Back to the Contents\)](#)

Taking the Exams

Personal Identification

On the date of your appointment, arrive at the testing center at least 15 minutes before the scheduled start time. You must bring two forms of identification with you. The first must be a current government-issued ID with your photograph and signature.

Examples of Acceptable Forms of Government Issued ID	Examples of Acceptable Forms of Supplemental ID	Examples of <u>Unacceptable</u> Forms of ID
• driver's license • passport • National identity card • military ID	• credit card • employee ID card	• Library card • Social Security card

Exam Format

The exams are computer-based. You will not receive any copies of the questions. Scrap paper can be used, but will be collected at the end of the exams.

Once admitted to the testing room, there will be a tutorial that introduces the functionality of the exam and a brief message from PRMIA. After this has been viewed, you may begin your exam. You will be asked multiple-choice questions, varying in quantity by exam.

Exam No.	Exam Name	No. of Questions	Time Allowed
I	Finance Theory, Financial Instruments and Markets	30	1.5 hours
II	Mathematical Foundations of Risk Measurement	24	2 hours
III	Risk Management Practices	36	1.5 hours
IV	Case Studies, PRMIA Standards of Best Practice, Conduct & Ethics	30	1 hour

If you take the **full exam**, it is broken into a 3 hour section, covering Exams I and II, then a 2 hour section for Exams III and IV. There is a break of one hour after the first section. Once you have started your break you will not be able to return to those questions. Following the break, you will have two hours to complete the remaining two sections of the exam. Allowed times include the time spent on the tutorial, thus the lesser time allocated when full exams are scheduled.

The testing system allows you to mark and review questions as long as time is remaining. Please note that you are unlikely to finish your exams with substantial extra time. You are encouraged to use the tutorial in an expedient manner, but sufficient time is allocated for you to complete the tutorial and each exam.

Calculators

An online scientific calculator is part of the testing computer. It is the same scientific calculator that is available with Windows operating systems. If you use the Windows

calculator on your computer, but it is not set to default to the scientific version, click View, then Scientific to change default formats. **We strongly recommend that you familiarize yourself with this calculator before arriving at the testing center.**

Arrive on time!

It is very important that you leave enough time to arrive at the testing center early. Candidates that arrive late to the test center may not be permitted to test. The full charge for the exam will be made if you are not admitted for any reason. As these centers offer exams for other organizations as well, not everyone in the room will be taking the same exam, so no assumptions should be made about when other candidates enter or leave the testing center.

Each testing center has an administrator who can assist candidates with any questions that they may have. All exam rooms are videotaped and monitored via a parabolic mirror by a proctor.

Exam results

Your exam results should be available within 15 business days of your test date.

Cancellations/Changes

There is a US\$35 charge per exam to re-schedule or cancel. The full test fee is charged for any cancellations or changes made 5 business days or less before the scheduled exam date. You can request a change/cancellation form from support@prmia.org.

[\(Back to the Contents\)](#)

PRM Syllabus

The subject matter of the PRM program is broken down broadly as:

Exam	Topic	Weighting
I	Finance Theory, Financial Instruments and Markets	25%
II	Mathematical Foundations of Risk Measurement	20%
III	Risk Management Practices	30%
IV	Case Studies, PRMIA Standards of Best Practice, Conduct and Ethics	25%

The complete syllabus is outlined below, including the weighting of each section.

Exam I – Finance Theory, Financial Instruments and Markets

Finance theory	<i>40%</i>
Portfolio theory and asset pricing	
<ul style="list-style-type: none">• Mean-Variance analysis• CAPM, APT, ICAPM, CCAPM• Efficient frontiers, capital market line, beta• Tobin approach	
Contingent claims	
<ul style="list-style-type: none">• The Black-Scholes-Merton model• The Binomial model• Put-call parity• Interest rate parity• Cash-and-carry pricing	
Financial instruments (descriptive & pricing knowledge)	<i>20%</i>
<ul style="list-style-type: none">• Compounding methods (simple, annual, semi-annual, continuous)• Simple (non-optional) bonds• Floating rate notes• Futures, Forwards• Swaps• Options• Basics of term structure modeling• Hybrid instruments• Convertible bonds• Caps, floors, swaptions• Simple exotics (such as barrier options)	
Markets	<i>40%</i>
<ul style="list-style-type: none">• Money market / FX market• Markets for commodities (natural gas and oil)• Capital markets	

Exam II – Mathematical Foundations of Risk Measurement

Calculus

25%

- Ordinary and partial derivatives
- Taylor series expansions
- Rate of change
- Optimization
- Area/volume
- Integration

Linear algebra

25%

- Matrix algebra (determinants, singular matrices, etc.)
- Positive definiteness
- Eigenvectors and Eigenvalues
- Cholesky factorization

Probability

50%

- Random variables
- General theory for univariate probability densities plus knowledge of standard distributions (uniform, normal, lognormal, Poisson, Chi-squared etc, conditional probabilities)
- Moments (up to fourth)
- Covariance and correlation matrices
- Principal component analysis
- Monte Carlo simulation
- Linear regression
- Basic statistical tests
- Coping with missing data

Exam III – Risk Management Practices

Market risk

33%

- Duration and convexity
- Cash flow maps, and PVBP interest rate sensitivity
- Greeks of instruments and portfolios
- Implied volatility and smile, smirk
- Value-at-Risk (VaR)
- Calculation of VaR for linear portfolios
- Monte Carlo and Historical calculation of VaR
- Covariance matrix construction (UWMA and EWMA)
- Market risk limits (stop-loss, exposure, VaR)
- Stress testing
- Scenario analysis
- Alternative risk measures
- RAROC & economic capital allocation

Credit risk*50%*

- Exposure, loss given default and expected loss
- Marginal vs cumulative default risk
- Corporate vs sovereign risk
- Settlement risk and netting systems
- Market prices, accounting (Altman) and actuarial methods
- Rating agencies and their grades
- Transition matrix
- Joint transition matrices and correlated migrations
- Credit derivatives (default swaps and total return swaps)
- Credit scoring models
- Recovery rate distributions
- Implied (from credit spread) default probability
- Merton model and KMV
- RAROC & economic capital allocation

Operational risk*17%*

- Typologies of operational risk
- Insurance, reinsurance
- Causal models
- Risk management processes, prevention and mitigation
- Loss events databases and their uses
- RAROC & economic capital allocation

Exam IV – Case Studies, PRMIA Standards of Best Practice, Conduct and Ethics**Case Studies***80%*

- Barings
- Metallgesellschaft
- LTCM
- Group of 30 Report

PRMIA Standards of Best Practice Conduct and Ethics, Bylaws*20%*

[\(Back to the Contents\)](#)

How to Prepare for the Exams

Practicing risk managers are already preparing for the exams by going to work each day where they interact with other professionals and read industry magazines, software manuals, company policies and procedures, academic journals, websites and regulatory notices. They attend risk committee meetings, prepare risk reports, give presentations and write papers. They attend PRMIA chapter meetings, participate in on-line forums and attend risk management conferences. All of these activities prepare candidates for the PRM exams.

Recommended Reading

The [Professional Risk Managers' Handbook: A Comprehensive Guide to Current Theory and Best Practices](#) is the recommended reading material for candidates preparing for the exams of the PRM certification program.

This Handbook covers all of the syllabus items for Exams I, II and III, while free web-based resources are available at [Reading List - Web-based Resources](#) for Exam IV.

Special pricing is available for registered candidates and the Handbook can be purchased at http://www.prmia.org/PRM_Handbook/HBPurchase.php.

You may also choose to purchase books published by other sources. We provide links to all of these resources below:

[Reading List - Books at Amazon.com](#)

PRM Diagnostic Exam

You may need to brush up on your risk management theory or on those areas you might not yet have worked on. PRMIA has developed an online [PRM Diagnostic Exam](#), scheduled for launch in November of 2003. The Diagnostic exam can be used to assess your current skills as they relate to the PRM Syllabus and/or as a practice exam prior to taking the PRM Exam(s)

Alternatively, you can do a self-assessment by looking at the exam content outlines to identify the topics you may need to study. We suggest the following resources to prepare for the exams:

PRM Self-Study Guide

This is a detailed study guide, produced by experts in the field, which encourages a structured learning approach to PRM exam preparation, whether alone or in self-study groups. The guide is free of charge and available to [download](#) from the PRMIA site.

The guide offers advice and motivation on examination techniques including a considered approach to modular multiple-choice timed exams. It provides considerable assistance in studying for the examinations, especially in the timing and ordering of study, key facts of knowledge, the ease of obtaining the study text reference and practice with worked examples.

Online PRM Training Courses

[Online training courses](#) designed specially for the PRM exams are accessible through the PRMIA web site. These courses have been developed and are administered by third parties such as KESDEE and Passpro. You may find these to be very helpful supplements to your study program, or substitutes for reading materials that are difficult to obtain in your location. You can learn more about these courses by [clicking here](#).

Study Groups

PRMIA offers four Online Study Groups that run for consecutive 13-week sessions. These Study Groups help guide you to completion of the requirements for the PRM designation in one year or less and let you connect with others around the world who are preparing at the same time as you. There is no cost to join the Online Study Groups. To join the group(s) of your choice [click here](#).

In addition, exam candidates in various cities have formed informal [study groups](#) to help one another prepare for the exam. Details vary from city to city, however the groups tend to meet weekly and participation is free and open to all. For information on joining a study group or setting one up for your staff, please contact support@prmia.org.

PRM Training Courses

Various training companies are offering [review courses](#) for the PRM program. They have submitted their programs to our Education Committee to ensure that the subject matter matches the PRM syllabus.

PRMIA Chapter Meetings

PRMIA has regional chapters around the world that host regular meetings on current risk management topics. The speakers are leading industry figures from banks, exchanges, regulators, academia, consultants, vendors, asset managers etc. The meetings are free and open to all PRMIA members, and constitute an excellent opportunity for professionals to stay informed of the latest developments in risk management and measurement.

[\(Back to the Contents\)](#)

FAQ

What is required to receive the PRM Designation? Candidates must demonstrate sufficient knowledge and understanding of the building blocks essential for the successful practice of modern risk management by passing exams in four specific areas. In addition, they must commit to further uphold the highest professional and ethical standards as defined by the [PRMIA Standards of Best Practice, Conduct and Ethics](#).

I've looked at some other designations in risk management, what are the reasons why so many people are choosing the PRM? Here are ten reasons why:

- Reason #1** Global - Offered in over 140 countries at nearly 4,000 testing centers, the PRM is a true international benchmark.
- Reason #2** Flexible - Risk managers' schedules are far from predictable, so you can take PRM exams any business day of the year and in any order you wish.
- Reason #3** Predictive Power - Because the PRM is broken into four exams, each of which must be passed to attain the designation, you cannot use strength in one area, say Math or Finance, to cover weaknesses in other area, say Risk Management. Exams with one overall score cannot reliably validate each critical area of competence. This muddling creates uncertainty about whether the candidate does indeed have the broad knowledge and understanding that risk managers must bring to their job. The point of a certification program is to remove such uncertainty.
- Reason #4** Affordable - This is also part of the global aspect. We recognize that not all economies are alike, so those who earn less than US\$25,000 per year are eligible for fee discounts.
- Reason #5** The PRM Handbook - Written by over 40 leading authors, the PRM Handbook is available anywhere in the world that has access to the Internet.
- Reason #6** Endorsements - Many companies have put their brand names on the line by publicly endorsing the PRM exam. There are no such firms doing this publicly for any other risk management certification program that we know of.
- Reason #7** Recognition of Other Achievements - We provide partial credit to those that have demonstrated their skills through attaining designations like the CFA, FSA, ASA and others.

Reason #8 No Maintenance Fees - Just because you are successful doesn't mean that PRMIA should earn an annuity. We will provide you with benefits as a successful PRM like training course discounts and access to resources that are only for PRMs. You won't need to pay us anything to keep your designation active.

Reason #9 Quality - Because the PRM is delivered via computer, in controlled testing facilities, we are able to monitor questions so that they truly reflect our syllabus and desired degree of difficulty. In other words, our exam is designed, not just assembled.

Reason #10 Respect - The PRM is the most challenging certification program for financial risk managers. Holders of the PRM are easily distinguished for their achievement and have the backing of PRMIA, the premier meeting place for the risk profession.

Between family and work, I don't have time to prepare for lengthy exams. Can you help me? Of course. The PRM program is first and foremost designed to help you and the profession to advance. That's our main goal. So, to help you with your busy schedule, you can take the exams required at your pace, in any order, over a period of up to two years. Individual exams vary in length from one to two hours, making them very easy to fit into most business days. If you find that you want to take two or three exams at once and save the one that you feel will be the most difficult for another time, you can do that. If you wish to take the exams in two different sittings, you can do that too. It's very easy to make time for the program.

What are the exam dates? You may take the exams on any business day. You may register for the date or dates that are most convenient for you.

What are the fees? The program allows for you to take one, two, three or all four exams at the same time. There are discounts for taking more than one exam at a time. Payment by credit card is required.

	2004
One Exam	US\$150
Two Exams	US\$235
Three Exams	US\$295
Full Program	US\$345

Do you offer scholarships for students and those of lower income? Limited scholarships are available to offset a part of the cost of the PRM program for anyone with annual income below US\$25,000 who will not have the costs of the program reimbursed by their employer. If you qualify, contact support@prmia.org to inquire about this program.

Are group discounts available? Discounts are also available for companies or institutions that wish to enroll 10 or more candidates in any part of the program. The candidates need not sit at the same time, and your group will have up to one year to use the testing vouchers purchased. Contact support@prmia.org to learn more.

Where does the money go? All of the net revenues from the exam program go to support education and resources for PRMIA members. In 2003, over 10,000 attendees are expected at free PRMIA chapter meetings along with over 400,000 unique visits to the PRMIA web site. So, your support of the PRM program gives you and your colleagues continued benefits. PRMIA is a tax-exempt, nonprofit professional association and does not have any for-profit subsidiaries that may benefit private individuals.

Are there any prerequisites? Membership in PRMIA, which is free, is the only prerequisite. [Click here](#) to become a member.

What is the format of the exams and how long are they? The exams are all multiple choice questions. PRMIA's computer-based method of delivery of the exams allows us to evaluate very specific details about each question administered, including down-to-the-second measurements of how long it takes most candidates to answer each question. Exams vary in length based on these validated testing times:

Exam No.	Exam Name	No. of Questions	Available Time
I	Finance Theory, Financial Instruments and Markets	30	1.5 hours
II	Mathematical Foundations of Risk Measurement	24	2 hours
III	Risk Management Practices	36	1.5 hours
IV	Case Studies, PRMIA Standards of Best Practice, Conduct & Ethics	30	1 hour
	Complete PRM Exam	120	6 hours

I just want to learn, not to be certified. Is this still for me? Yes. There is no requirement that you complete all exams. If you want to improve your knowledge and understanding, or want to train your staff, pick and choose the areas that you think they should know and use our program to validate their success. All financial market participants benefit from a better understanding of risk management Case Studies. You might have your trading desk sit for Exam IV. Suppose someone is applying for an entry level job in your department. You can have them take Exam I or Exam II to determine how ready they are to become a risk manager. The goal of our program is to advance you and the profession.

How do I register? Registration is available online and via fax. Go to <http://www.prmia.org/certification/register.php> for details. Spaces at the testing centers will be reserved on a first-come, first-served basis, so you are encouraged to register as soon as possible.

What is the expected pass rate? Pass rates vary by exam and tend to be in the high 50's to high 60's in percentage terms. Overall, the success rate of candidates attempting to attain the PRM has been just above 50%.

Where can I take the exams? The exams are offered in all over the world in secure, clean and convenient testing centers. [Click here](#) to find specific addresses of testing centers.

What are the security arrangements? Candidates will be required to present two forms of official identification (see detailed policy in page 6) when arriving at the test center. Candidates must sign-in and sign-out of the testing center. No papers may be brought into the examination room. Many rooms are videotaped during testing and all are monitored by a proctor using a parabolic mirror or other viewing device. Scratch paper is provided at the site, but candidates must leave all papers in the testing room upon completion of the exam.

Can I bring a calculator or PDA into the exam? You will not be allowed to bring anything into the testing room.

What Designation is awarded to successful candidates? With the awarding of the certificate, successful candidates become a Professional Risk Manager, or PRM, and will have a limited license to use the PRM insignia on business cards, resumes and CV's.



Do you recognize other Designations? Holders of the CFA, CQF, ChIA, Actuarial Fellow or Actuarial Associate designations will receive partial credit towards the PRM designation. The credit applied is based on the quality and challenge of materials covered in each of those programs, the rigor of the testing versus that required for the PRM designation and the ability to interpret results as being equivalent to the requirements of the PRM. When registering, select the appropriate "Cross-Over" exam for the designation that you hold. FRMs from 1997 - 2001 test dates should contact support@prmia.org for special assistance before registering.

If I am successful, do I have to pay an annual fee to keep my designation? No. Some organizations want you to pay them for your achievement. PRMIA wants to encourage you to develop professionally and to attain the highest standard in the industry. PRMIA will give to you special discounts and special resources as a PRM.

Who has endorsed the PRM? The PRM program has received the [endorsements](#) of some of the leading firms in our business because it is simply a better test of skills and it is available to members in all parts of the world. No other program has received such public backing. [View a PowerPoint](#) about the PRM benefits, or read the [Testimonials and Case Studies](#) of some successful PRMs.

I'm ready to register, what should I do? Go to <http://www.prmia.org/certification/register.php> and book your place!

[\(Back to the Contents\)](#)

Sample Exam Questions

The following sample questions should give you a flavor for the format and content of the actual exams. They are only part of the length of the actual exams and therefore do not cover all subjects contained in the detailed content description provided in this document. Questions on any of the subjects listed previously may appear on the actual exam.

EXAM I FINANCE THEORY, FINANCIAL INSTRUMENTS AND MARKETS

1. Assume you live in a CAPM world and the expected return on the market portfolio is 9%, while the risk free rate is 3%. If the beta of stock A is 1.3, the expected return on A is:

- a) 14.7%
- b) 12.9%
- c) 10.8%
- d) 16.8%

2. In the valuation of derivatives, the expression "change of measure" means:

- setting the drift to zero
- change of volatility
- both (a) and (b)
- none of the above

3. Which of the following is true?

- a) Non-Markovian interest rate processes are usually represented by recombining trees
- b) Markovian interest rate processes are usually represented by recombining trees
- c) Non-Markovian interest rate processes are usually represented by trinomial trees
- d) none of the above

4. Which of the following is/are true concerning preferred stocks?

They are somehow similar to subordinated debt, but unlike bond holders preferred share holders could not force a company into bankruptcy if preferred coupons (dividends) were not paid on time

- a) Many preferred shares provide for cumulative preferred dividend payments having priority over ordinary dividends
- b) From an issuer's tax perspective, preferred stocks are a more expensive source of financing than bonds
- c) All of the above

5. A fixed income instrument will pay 12-month Libor on a 1,000 Swiss Francs (CHF) face value two times: one year from today and two years from today (no principal payment). The rates are set in arrears (payments at the end of a year reflect the Libor rate at the beginning of the year). What is the price of this instrument if the (zero-coupon) two-year CHF swap curve is a 3% for all maturities?
- a) CHF 57.4
 - b) CHF 1000
 - c) CHF 1067
 - d) CHF 67.6
6. In a long option straddle strategy, where one buys a put and a call simultaneously at the same strike, the following is true:
- a) Delta will be zero, regardless of the level of the spot price
 - b) Gamma will be the highest at the money and approaching maturity
 - c) Delta will be near to 1 at the money and approaching maturity
 - d) Gamma will be zero at the money and approaching maturity
7. Herstatt Risk relates to:
- a) the market risk of an FX contract
 - b) the German Mark debacle of 1978
 - c) the settlement risk of an FX contract
 - d) none of the above
8. If a dealer sells a security to a retail customer and the customer pledges to resell that security to the dealer, the trade is called:
- a) a reverse repo
 - b) a forward sale
 - c) a repo
 - d) a reverse forward
9. A gas market maker (MM) has agreed to deliver gas at \$3/MMBtu 6 months from now. The spot price for gas is \$2.50/MMBtu, the 6 month forward price is \$2.75/MMBtu, the interest rate is 6% and storage cost is \$0.03/month per MMBtu. The MM is confident that the price would be \$2.70 six months from now. Given the MM's market view, what is the best strategy for the MM to meet its obligation?
- a) Buy on the spot market
 - b) Buy the forward
 - c) Do nothing
 - d) MM is indifferent between c) and b)

10. Assume the following price curve for crude oil in bbl and ignore the time value of money.

Month	Price
1	\$21
2	\$22
3	\$23
4	\$24
5	\$25
6	\$26

A customer wants a tailored six month swap with constant volumes but request the fixed price for the last two months to be set at \$20/bbl. What must be the fixed price for the first four months?

- a) not determinable
- b) 26.0
- c) 23.7
- d) none of the above

EXAM II MATHEMATIC FOUNDATIONS OF RISK MEASUREMENT

1. Find a linear polynomial $p(x)$ that is a tangent-line approximation for the function:

$$f(x) = e^{2x-4}$$

at the point $x_0 = 3$.

- a) $14.778x - 36.945$
 - b) $7.389x + 2.718$
 - c) $2.718x$
 - d) $14.778x + 0.018$
2. Evaluate the definite integral:

$$\int_0^2 xe^{x^2} dx$$

- a) 5.437
 - b) 26.799
 - c) 21.285
 - d) 7.389
3. Which of the following statements is false?
- a) Singular matrices have determinant 0.
 - b) Singular matrices have columns that are not independent vectors.
 - c) A product of two non-singular matrices can be singular.
 - d) Singular matrices have 0 as an eigenvalue.

4. Determine the inverse matrix of:

$$\begin{pmatrix} 1 & 2 \\ 0 & 2 \end{pmatrix}$$

a) $\begin{pmatrix} 1 & 1 \\ 0 & 0.5 \end{pmatrix}$

b) $\begin{pmatrix} 1 & -1 \\ 0 & 0.5 \end{pmatrix}$

c) $\begin{pmatrix} 0.5 & -1 \\ 0 & 1 \end{pmatrix}$

d) $\begin{pmatrix} 1 & -1 \\ 0.5 & 0 \end{pmatrix}$

5. What can we say about the sum $X + Y$ of two independent normal random variables X and Y :

- a) It is normal only if X and Y have the same mean.
- b) It is always normal.
- c) It is chi-squared.
- d) It is chi-squared if X and Y both have mean 0.

6. What is the formula for the skewness of a random variable X that has mean μ and standard deviation σ ?

a) $\frac{E([X - \sigma]^2)}{\mu^2}$

b) $\frac{E([X - \mu]^4)}{\sigma^4}$

c) $\frac{E([X - \mu]^3)}{\sigma^3}$

d) $\frac{E([X - \mu]^4)}{E([X - \sigma]^4)}$

7. A covariance matrix for a random vector:

- a) Is strictly positive definite, if it exists
- b) Is nonsingular, if it exists
- c) Always exists
- d) None of the above

8. What is the standard deviation of a random variable Q with probability function:

$$\phi(q) = \begin{cases} .25 & q = 0 \\ .25 & q = 1 \\ .50 & q = 2 \end{cases}$$

- a) .6875
- b) .4727
- c) .8291
- d) .4281

EXAM III RISK MANAGEMENT PRACTICES

1. Under the standard parametric VaR methodology, which of the following assumptions is true?

- a) Returns follow a log normal distribution
- b) Log returns follow a normal distribution
- c) Mean log return is zero for daily VaR
- d) All of the above

2. Under the RiskMetrics cashflow mapping method for interest rates, price volatility is required. The formula to convert yield volatility (expressed as a percentage of current yield) to price volatility is:

- a) Price Vol (σ_p) = Modified Duration (MD) \times Interest Rate (Y) \times Yield Vol (σ_y)
- b) Price Vol (σ_p) = MD \times $\frac{\sigma_y}{Y}$
- c) Price Vol (σ_p) = MD \times σ_y
- d) Price Vol (σ_p) = MD \times (1 + Y) \times σ_y

3. For EWMA (Exponentially Weighted Moving Average), using a decay factor of 0.94 and a tolerance level of 1% (i.e. excluding exponential weights below 1%), the effective number of data points used to estimate the covariance matrix is:

- a) 74
- b) 150
- c) 100
- d) 250

4. The portfolio has one risky bond from company A. Company A is a subsidiary of XYZ and if XYZ defaults Company A does so too. The probability of default of XYZ is 0.3 and the probability of company A going into bankruptcy without XYZ defaulting is 0.5. What is the probability of having a default on the risky bond?

- a) Cannot be determined
- b) 0.60
- c) 0.70
- d) None of the above

5. Assuming independence and a recovery rate of 70%, what is the expected loss on the following portfolio?

	Face value of the bond	Probability of default
Bond A	1,000 Euros (EUR)	0.4
Bond B	2,000 EUR	0.3

- a) 300 EUR
 b) 900 EUR
 c) 1,000 EUR
 d) None of the above
6. How is the loss given default incorporated in the CreditMetrics Technical Document?
- a) The document does not consider it
 By a parameterized distribution
 b) By a look up table
 c) By a constant
7. What is characteristic of a default mode credit risk model?
- a) A model which considers two states of nature
 b) A model which incorporates a default definition
 c) A model which considers default as a reflecting state
 d) None of the above
8. What is the one-year transition matrix assuming only two categories [*i.e.* default (d) and non-default (nd)] from a portfolio with 300 loans and the following payment history?

Number of loans that transited from non default to default in 2000										
Jan	Feb	March	Apr	Jun	Jul	Aug	Sep	Oct	Nov	Dec
0	9	0	0	6	0	3	3	0	6	3

- a)

.7	.3
0	1
- b)

.9	.1
0	1
- c)

.9	.1
.1	.9
- d)

.3	.7
1	0

9. The Merton (1974) model implies that a position in a credit-sensitive bond is equivalent to:
- a) A long position in the firm's equity and a short position in a risk-free bond
 - b) A long put and a long call position on the firm's assets
 - c) A long position in a credit-risk-free bond and a short put on the firm's assets
 - d) An up-and-in call on a credit-risk-free bond and a short call on the firm's equity
10. Given a one-year probability of default of 20%, what would be the cumulative probability of default for the bond for the three years?
- a) 45.4%
 - b) 48.8%
 - c) 60.5%
 - d) None of the above
11. The Bank for International Settlement's, Basel Committee on Banking Supervision, has defined Operational risk as "The risk of loss due to inadequate or failed internal processes, people, and systems, or from external events". This definition excludes:
- a) Reputational risk
 - b) Strategic risk
 - c) Legal risk
 - d) a) and b)
12. The Bank for International Settlement's, Basel Committee on Banking Supervision recommends the operational risk management process at the corporate and business unit levels to be validated by:
- a) Audit
 - b) A committee of the board of directors
 - c) A designated member of senior management
 - d) None of the above

**EXAM IV CASE STUDIES, PRMIA STANDARDS OF BEST PRACTICE,
CONDUCT AND ETHICS**

1. Which position would have partially hedged Nick Leeson's primary option position at Barings?
- a) Long futures
 - b) Short Strangle
 - c) Long Strangle
 - d) Total Return Swap

2. Nick Leeson tried to hide his losses using what method?
 - a) Portaling
 - b) Switching
 - c) Re-margining
 - d) Volatility Smiles

3. What caused the losses for Metallgesellschaft?
 - a) At the final maturity date the price in the futures was well below the market price
 - b) To hold the position, they assumed a constant interest rate to invest the proceeds
 - c) At the final maturity date the price in the futures was well above the market price
 - d) To hold the position, they assumed an unbounded pool of resources

4. In the Metallgesellschaft case, what was the communication problem between the subsidiary and the parent company?
 - a) They did not communicate the loss as soon as it started to kick in
 - b) They did not communicate the intrinsic bet on the price of gas
 - c) They did not explain the economics of the strategy
 - d) They did not explain why they had to buy more future contracts

5. The strategy of getting creditors to lend money and invest equity in LTCM had the effect of:
 - a) Increasing transparency to the creditors
 - b) Reducing leverage
 - c) Giving LTCM partners a put on the value of the fund
 - d) No relevant effect as equity and debt offset each other

6. LTCM's balance sheet as of August 31, 1998 showed the following:
 - a) \$100 billion in assets, \$-0.5 billion in equity
 - b) \$125 billion in assets, \$2.3 billion in equity
 - c) \$400 billion in assets, \$4.0 billion in equity
 - d) \$125 billion in assets, \$6.1 billion in equity

7. According to the G-30, derivative credit exposure should be measured by:
 - a) Current Exposure
 - b) Potential Exposure
 - c) a) plus b)
 - d) a) plus b) minus Posted Collateral

8. According to the G-30 report, an ISDA master agreement is:
- a) Sufficient to prevent loss from counterparty default
 - b) Not substantially enhanced by a netting provision as bankruptcy courts widely recognize netting as a best practice
 - c) Enhanced when multiple master agreements exist between the same counterparties so that the legal risk of an oversight in documentation is reduced
 - d) None of these
9. Which of the following are excluded from being Regular Members of PRMIA by the PRMIA Bylaws?
- a) Corporations
 - b) Corporations and former Risk Managers
 - c) Regulators
 - a) Students
10. Which of the following is not part of PRMIA's guidance on Best Practices?
- a) Only standard methods of assessing risk should be used
 - b) PRMIA members must possess, be under the supervision of someone who possesses, or inform their supervisor of the lack of required skills and/or certification to complete their risk assessment work.
 - c) PRMIA members must not intentionally deceive others
 - d) PRMIA members must value validation of their work by peers

ANSWERS

Exam I	Exam II	Exam III	Exam IV
1 c	1 a	1 c	1 c
2 d	2 b	2 a	2 b
3 b	3 c	3 a	3 d
4 d	4 b	4 d	4 c
5 a	5 b	5 a	5 c
6 b	6 c	6 b	6 b
7 c	7 d	7 a	7 d
8 c	8 c	8 b	8 d
9 c		9 c	9 a
10 d		10 b	10 a
		11 d	
		12 a	

[\(Back to the Contents\)](#)