

 JUTA

Statistical Methods and Calculation Skills

Isabel Willemse

Third edition



Statistical Methods and Calculation Skills

This One



WBB1-DCQ-55KU

Copyrighted material

Statistical Methods and Calculation Skills

Isabel Willemse



Statistical Methods and Calculation Skills

First published 2001

Second edition 2003

Third edition 2009

© Juta & Co Ltd, 2009

PO Box 14373, Lansdowne, 7779, Cape Town

ISBN 978-0-70217-753-8

Disclaimer

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage or retrieval system, without prior permission in writing from the publisher. Subject to any applicable licensing terms and conditions in the case of electronically supplied publications, a person may engage in fair dealing with a copy of this publication for his or her personal or private use, or his or her research or private study. See Section 12(1)(a) of the Copyright Act 98 of 1978.

Project management: Sharon Steyn

Editing: Craig Farham

Proofreading: Lee-Ann Ashcroft

Cover design: Marius Roux

Design and typesetting by Mckore Graphics

Illustrations by Mckore Graphics

Printed and bound in the Republic of South Africa by Shumani Printers

The authors and the publisher have made every effort to obtain permission for and to acknowledge the use of copyright material. Should any infringement of copyright have occurred, please contact the publisher, and every effort will be made to rectify omissions or errors in the event of a reprint or new edition.

Contents

PART ONE: STATISTICAL METHODS	1
Unit 1: Introduction	3
1.1 Problem-solving steps	3
1.2 Definition	5
1.3 The language of statistics	5
1.4 Measurement	6
1.5 Role of the computer in statistics	8
Test yourself 1	9
Unit 2: Collection of data	12
2.1 Sources of data: where to get the data	12
2.2 Primary data sources	13
2.3 Questionnaire design	15
2.4 Selecting a sample	17
2.5 Non-random sampling	20
2.6 Random sampling	20
Test yourself 2	23
Unit 3: Summarising data using tables and graphs	25
3.1 Summarising qualitative data in tables and graphs	26
3.2 Summarising quantitative data in tables	36
3.3 Summarising quantitative data using graphs	46
3.4 Using software	51
3.5 Using visual aids	52
Test yourself 3	52
Unit 4: Summarising data using numerical descriptors	60
4.1 Measures of central tendency	61
4.2 Measures of dispersion	73
4.3 Measures of shape	82
4.4 Interpreting centre and variability	85
4.5 Measures of relative standing	89
4.6 Measuring dispersion using measures of relative standing	92
Test yourself 4	97

Unit 5: Index numbers	102
5.1 Construction of a simple index number	103
5.2 Construction of composite (or aggregate) index numbers	104
5.3 Additional topics on index numbers	109
Test yourself 5	111
Unit 6: Summarising bivariate data: regression and correlation analysis	114
6.1 Response variable (y) and explanatory variable (x)	115
6.2 Scatter diagram	115
6.3 Correlation analysis (r)	118
6.4 Regression analysis	121
6.5 Spearman rank correlation coefficient (r_s)	124
Test yourself 6	126
Unit 7: Time series	129
7.1 Components of a time series	129
7.2 Histogram	131
7.3 Time-series decomposition	131
Test yourself 7	144
Unit 8: Probability: basic concepts	148
8.1 Language of probability	149
8.2 Approaches to assigning probabilities	149
8.3 Properties of probabilities	152
8.4 Forming new events	154
8.5 Probability rules for compound events	156
8.6 Counting the possibilities	164
Test yourself 8	166
Unit 9: Probability distributions	173
9.1 Discrete probability distributions	173
9.2 Probability distributions for continuous random variables	178
Test yourself 9	184
Unit 10: Statistical inference: estimation	187
10.1 Statistics and parameters	187
10.2 Sampling distribution of the means	187
10.3 Estimating population parameters	189
10.4 Sample size	194
Test yourself 10	196
Unit 11: Hypothesis testing	198
11.1 Steps to follow in a single sample hypothesis test	199
11.2 Testing the difference among means and proportions	206
11.3 Tests using the chi-square distribution (χ^2)	209
Test yourself 11	215

PART TWO: CALCULATION SKILLS	221
Unit 12: Elementary calculations	223
12.1 The electronic calculator.....	223
12.2 The number system.....	224
12.3 Common notation.....	227
12.4 Basic operations.....	228
12.5 Signed numbers.....	231
12.6 Exponents (powers) (x^y).....	231
12.7 Square roots ($\sqrt{\quad}$).....	232
12.8 Logarithms (log).....	232
12.9 Factorial notation (!).....	233
12.10 Sigma notation (Σ).....	233
12.11 Fractions.....	234
12.12 Decimal numbers.....	236
12.13 Scientific notation.....	237
12.14 Rounding off decimals.....	238
12.15 Significant digits.....	239
12.16 The metric system.....	242
Unit 13: Percentages and ratios	243
13.1 Percentage calculations.....	243
13.2 Ratio (proportion) calculation.....	247
13.3 Business applications.....	247
Test yourself 13.....	251
Unit 14: Equations and graph construction	253
14.1 Graph construction.....	253
14.2 Solution of equations.....	254
Test yourself 14.....	257
Unit 15: Interest calculations	259
15.1 Basic concepts.....	259
15.2 Simple interest.....	259
15.3 Compound interest.....	261
15.4 Nominal and effective rates of interest.....	263
15.5 Annuities.....	263
Test yourself 15.....	267
Appendix 1: The standard normal distribution	269
Appendix 2: The t-distribution	270
Appendix 3: The chi-square distribution	271
Appendix 4: Random numbers	272

PART

1

Statistical Methods

UNIT 1

Introduction

This unit deals with the role of statistics in the data analysis process. Concepts that are basic to the study of statistics are discussed.

After completion of this unit you will be able to:

- recognise the role of statistics in life
- understand the language of statistics
- select suitable measuring scales for different types of data
- understand the role of computers in statistics.

We live in an era where we are faced with increasing amounts of information, also referred to as data. Every time you read a magazine or newspaper, listen to a news bulletin or advertisement, you see statistics. People quote numbers or statistics to support whatever it is they wish you to believe. Therefore, to perform many tasks efficiently in today's world, you need to have a basic understanding of statistical methods.

By itself, data cannot tell you much. When collected and used properly, data and the statistics calculated from them can help you to understand situations in order to evaluate your options and make informed decisions.

To be an informed consumer of information, you must be able to:

- extract information from tables and graphs
- follow numerical arguments
- understand the basics of how data should be gathered, summarised and analysed to draw statistical conclusions.

1.1 Problem-solving steps

1. Understand the problem or question you hope to answer.
2. Identify an appropriate data source and decide how to measure it. The researcher must decide whether an existing data source is adequate or whether new data must be collected.
3. Determine if you will use an entire population or a representative sample. If using a sample, decide on a viable sampling method.
4. Collect the information needed to answer the problem.
5. Organise and summarise the information. Tables, graphs and numerical summaries allow increased understanding and provide an effective way to present data. This

initial analysis provides insight into important characteristics of the data and gives guidance in selecting appropriate methods for further analysis.

6. Analyse the data in order to draw conclusions, make recommendations and assess the risk of an incorrect decision. This usually involves generalising from a small group or sample of individuals or objects that we have studied to a much larger group or population.
7. Communicate the results.

Example 1.1

Little information exists on the effects that antihypertensive drugs have on patients who have heart disease and normal blood pressure. Blood pressure is the force of blood against the arteries and is presented as two numbers: the systolic pressure (as the heart beats) over the diastolic pressure (as the heart relaxes between beats). A blood pressure measurement of 120/80 mm Hg (millimetres of mercury) is normal. Hypertension or high blood pressure exists in individuals with a systolic blood pressure above 160 mm Hg or a diastolic blood pressure above 100mm Hg. The researchers wanted to determine the effectiveness of an antihypertensive drug (10mg amlodipine) on preventing cardiovascular events such as congestive heart failure, stroke, or other heart-related problems.

(Source: Journal of the American Medical Association, Vol. 292, No. 18).

1. **Identify the problem.**

To determine the effectiveness of the drug on preventing cardiovascular events in patients who have heart disease and normal blood pressure.

2. **Collect the information** needed to answer the question.

The researchers divided 1 317 patients with heart disease and diastolic blood pressure less than 100mm Hg into two groups. Group 1 had 663 patients and group 2 had 654 patients. The patients in group 1 received 10 mg daily of the antihypertensive drug. The patients in group 2 received a placebo. A placebo is an innocuous drug such as a sugar tablet. Group 1 is called the experimental group and group 2 the control group. Neither the doctor administering the drug nor the patient knew whether he or she was in the experimental or control group. After 24 months of treatment, each patient's blood pressure was recorded. In addition, the number of patients in each group who experienced a cardiovascular event was counted.

3. **Organise and summarise** the information.

Before administering any drugs, it was determined that both groups had similar blood pressure. After the 24-month period ended, the experimental group's blood pressure decreased by 4.8/2.5 mm Hg, whereas the control group's blood pressure increased 0.7/0.6 mm Hg. In addition, 16.6% of patients in the experimental group experienced a cardiovascular event, while 23.1% of patients in the control group experienced a cardiovascular event.

4. **Draw conclusions** from the data.

You can extend the results from the sample of 1317 patients to all individuals who have heart disease and normal blood pressure. That is, the antihypertensive drug appears to decrease blood pressure and seems effective in reducing the likelihood of experiencing a cardiovascular event such as a stroke.

Key components of statistical thinking:

- use data whenever possible to guide the analysis
- look for connections and relationships
- understand why data values differ from one another.

1.2 Definition

Statistics is the scientific discipline that provides methods to help us make sense of data by:

- collecting data in a methodical way
- organising and summarising data using tables, numbers and graphs
- analysing data to draw conclusions or to answer questions.

The field of statistics can be subdivided into descriptive statistics and inferential statistics.

Descriptive statistics includes the collection and summarising of data to give an overview of the information collected.

Inferential statistics is the process of making an estimate, prediction or decision about a population based on sample data. Because a population is almost always very large, a sample is drawn from the population of interest and summarised using descriptive techniques. These results are then used to make decisions about the population. Such conclusions are not always going to be correct and it is therefore necessary to measure the reliability using the **confidence level** and the **significance level**. The confidence level measures the proportion of times that an estimating procedure will be correct over the long run. When the purpose of statistical inference is to draw conclusions about the population, the significance level measures how frequently the conclusion will be wrong. A 2% significance level means that in the long run, this type of conclusion will be wrong 2% of the time.

1.3 The language of statistics

An **investigation** or **experiment** is any process of observation or measurement.

Elements are the people or objects about which information is collected.

A **population** is a complete collection of elements you wish to study. If the population contains a countable number of items, it is said to be **finite**, and when the number of items is unlimited, it is said to be **infinite**. A study of the entire population is known as a **census**. A **parameter** is a numerical measure that describes the population. It is calculated using all the data of the population, such as an average. It is usually indicated by a letter from the Greek alphabet (e.g. μ , σ , π).

A **sample** is a portion of data drawn from the population. The sample must be representative of the population. A **statistic** is a numerical measure that describes a sample. It is usually indicated by a letter from the Roman alphabet (e.g. \bar{x} , s , n , p).

A **variable** is a characteristic of interest about each element of a population or sample. It is the topic about which data are collected, such as the age of the first year students at the university or the weight of each first year student. Not all students have the same age or weigh the same; this will vary from student to student. That means there is a variation in the weights or ages. If there were no variability in the weights or

ages, statistical inference would not be necessary. The observed values of the variable are the **data** we will use in a statistical investigation.

Variables can be classified as quantitative or qualitative.

Quantitative variables provide numerical measurements of the elements of the study. Arithmetic operations such as addition and subtraction can be performed on the values of the variable.

Qualitative or categorical variables provide information that is non-numerical, like marital status, type of job, gender, etc. Qualitative information can sometimes be coded to make it appear quantitative but will have no meaning on a number line.

We can further classify quantitative variables as discrete or continuous.

Discrete variables are countable and can assume a countable number of values, such as the number of potatoes on the plant. Fractional values can also occur, but must have distance between them, for example interest rates and stock prices.

If you measure to get the value of the variable, it is **continuous**. It has an infinite number of possible values that are not countable. For example weight, length, time taken to complete a task, age, etc. can be measured to any desired accuracy or number of decimal places within a given range.

Example 1.2

Distinguish between qualitative and quantitative variables:

1. Gender: it is a qualitative variable because it allows a researcher to categorise the individual as male or female. No arithmetic operations can be performed with this data.
2. Temperature: it is a quantitative variable because it is numeric and arithmetic operations such as addition and subtraction provide meaningful results.
3. Postal code: it is qualitative because it indicates a location. Although the code is in numbers, addition and subtraction of the codes does not provide meaningful results.
4. Number of drinks at a party for a couple of friends: it is quantitative because it provides numbers which can be used in arithmetic operations.

Example 1.3

Distinguish between discrete and continuous variables:

1. The number of heads obtained after flipping a coin five times: discrete because we can count the number of heads obtained.
 2. The number of cars that arrive at the KFC drive-through between 10h00 and 12h00: discrete because we can count the number of cars.
 3. The distances different model cars with the same tank capacity can drive in city driving conditions: continuous because we measure the distances.
 4. Temperature: continuous because we measure temperature.
-

1.4 Measurement

Measurement is the process we use to assign numbers to the observations or elements of a variable. The term number does not necessarily mean numbers that can be added,

subtracted, multiplied or divided. Instead, it means that numbers are used as symbols to represent certain characteristics like age, income, height of the object, person, etc. For example, as a student your student number may identify you.

There are four levels or scales of measurement, each with its own characteristic and from the weakest to the strongest; they are nominal, ordinal, interval and ratio. The analysis you carry out depends on the type of scale used to measure the characteristics of the variable.

1.4.1 Nominal scale

This level, also known as a categorical level, applies to data that consist of names, labels and categories in *no specific order*. Numbers or symbols are used to identify groups to which various observations belong. For example in counting males and females, the male group can be assigned the code 1 and the females the code 2. These nominally scaled numbers serve only as a label for the group and the measurement consists of placing the data in the correct group. No arithmetic operations can be performed by such numbers other than counting the groups and the number of elements falling into each group.

1.4.2 Ordinal scale

The categories into which objects are grouped are ranked in *some order* using numbers or symbols. Items can be classified not only as to whether they share some characteristic with another item, but also whether they have more or less of this characteristic. Differences between data values either cannot be determined or are meaningless. For example, income levels such as low, medium or high. The permissible analysis methods for ordinal data include techniques generally associated with the order of the observations.

1.4.3 Interval scale

This scale applies to data that can be arranged in order. In addition, differences between data values are meaningful but ratios of data are not. Temperature is a classic example of an interval scale: the increase on the centigrade scales between 10 and 20 is the same as the increase between 30 and 40. However, heat cannot be measured in absolute terms (0°C does not mean no heat) and it is not possible to say that 40° is twice as hot as 20° . Interval-level *data may not have an absolute zero starting point*. This sometimes causes difficulties in interpreting the interval-scale data. Arithmetic operations can be performed on the difference between numbers, not the numbers themselves.

The following are examples of data at the interval level of measurement:

- calendar dates
- time
- shoe sizes
- Celsius-scale temperatures.

1.4.4 Ratio scale

The ratio level of measurement applies to data that can be arranged in order. Both differences between data values and ratios of data values are meaningful because **a true zero exists**. Arithmetic operations can be performed on the numeric values themselves. Money is an example of the ratio scale of measurement: the zero point is meaningful – that is, at zero you have none; and R10 is twice as much as R5.

Activity 1.1

Categorise these measurements relating to fishing according to level:

1. species of fish in the Vaal dam
 2. cost of rod and reel
 3. time of return home
 4. rating of fishing area: poor, fair, good
 5. number of fish caught
 6. temperature of water.
-

1.5 Role of the computer in statistics

In all aspects of business life we are likely to encounter increasing quantities of data. Computers and new information technologies literally put data at our fingertips; for example stock levels in a warehouse some distance away, or share prices in Japan can be known in minutes.

The Internet can provide access to data across continents at low costs. The challenge is to organise and analyse this information, in such a way that managers can make sense of it by making use of statistical and quantitative techniques. Facilities like spreadsheets or statistical and mathematical software packages make such analysis readily available to everyone. The use of such computer software assumes that you are able to interpret the output that can be generated, not only in a strictly quantitative way, but also in terms of assessing its potential to help in business decision making.

Computers also provide the opportunity to experiment with and explore data in ways that would not otherwise be possible.

A computer may be efficiently used in any processing operation that has one or more of the following characteristics:

- large volume of input
- repetition of projects
- desired greater speed in processing
- greater accuracy
- processing complexities that require electronic help.

It can help you develop your ideas about how to organise the information by using a 'try and refine' approach, which can take too long to carry out manually.

TEST YOURSELF 1

1. A survey of 100 people is conducted and all are asked questions relating to the following characteristics:
- marital status
 - salary
 - occupation
 - number of hours of television they watch per week.

What type of data and measurement scales are applicable?

2. Human beings have one of four blood types: A, B, AB or O. What type of data do you receive when you are told your blood type?
3. The personnel manager of a business is studying employee morale and uses a questionnaire to collect data. A typical question on the questionnaire: 'I feel that I am performing a valuable service for society when I do my job well.' Circle the letter that most closely represents your agreement with the statement;

Strongly agree	Agree	Undecided	Disagree	Strongly disagree
A	B	C	D	E

For the data generated by this question state:

- a) the elements to be observed
 - b) the variable being measured
 - c) whether the data are quantitative or qualitative
 - d) the measurement scale that should be used to record the variable.
4. • "Every week a clerk in a hypermarket records the number of transactions that occurred that week at each of the checkout tills."
 • "Once an hour a random sample of 100 battery chargers is selected from an assembly line and the number of defective chargers is recorded."

For the above two statements:

- a) What elements are being observed?
 - b) Define the variable.
 - c) What type of data is being used?
 - d) What is the measurement scale of each data set?
5. Say whether each of the following variables is quantitative or qualitative and indicate the measurement scale that is appropriate for each:
- a) Age of a respondent to a consumer survey.
 - b) Sex of respondent to a consumer survey.
 - c) Thickness of the gelatine coating of a vitamin E capsule.
 - d) Make of motorcar owned by a sample of 50 drivers.
 - e) Percentage of people in favour of the death penalty in each of the provinces.
 - f) Concentration of a contaminant (microgram per cubic centimetre) in a water sample.
 - g) The amount by which a 1 kg package of beef mince decreases in weight because of moisture lost before purchase.
 - h) The length of a 1-year-old rattlesnake.

6. Based on a study of 2 050 children between the ages of two and four, researchers concluded that there was an association between iron deficiency and the length of time that a child is bottle-fed. Describe the sample and the population of interest for this study. Define the variables and type of data that were used.
7. The leader of a rural community is interested in the proportion of property owners who support the construction of a sewer system. Because it is too difficult to reach all 7 000 property owners, a survey of 500 owners, selected at random, is undertaken. Describe the population and sample for this problem. Define the variable of interest and type of data that will be needed.
8. The student council at a university with 10 000 students is interested in the proportion of students who favour a change in the grading system. Two hundred students are interviewed to determine their attitude toward this proposed change. What is the population of interest? What group of students constitutes the sample in this problem?
9. All South Africans are involved in at least one form of gardening. This result shows that gardening is the number one leisure activity. Classify this study as either descriptive or inferential.
10. A random sample of 200 academic staff members was taken at a university. Each was asked the following questions:
 - What is your rank (lecturer, senior lecturer, professor)?
 - What is your annual salary?
 - In which faculty (Business, Engineering, Arts) are you employed?
 - How many years have you been employed?Identify the type of data as quantitative or qualitative. If quantitative, classify it as either discrete or continuous. Indicate the measurement scale in each category.
11. For each of the following examples, determine the type of data and measurement scale:
 - a) the month of highest sales for each supermarket in a sample
 - b) the weekly closing price of gold throughout the year
 - c) country of origin
 - d) a taste tester's ranking (best, worst, etc.) of four brands of tomato sauce for a panel of 10 testers
 - e) the size of soft drink (small, medium, large) ordered by a sample of Big Burger customers
 - f) the marks achieved by the students in a statistics exam in which there were 5 questions, each worth 10 marks
 - g) the letter grades received by students in a statistics course (A, B, C, D, E)
 - h) do you have season tickets for Ellis Park?
 - i) would you rate the Lotto Show as excellent, good, fair or poor?
 - j) the length of service for several members of a hospital staff: Sara - 10 years; Rob - 20 years; Grace - 25 years; and Alfonds - 30 years
 - k) the number on a rugby player's jersey
 - l) grams of carbohydrates in a doughnut
 - m) number of unpopped kernels in a bag of microwave popcorn
 - n) volume of water lost each day through a leaky faucet.

12. For each of the following case studies, identify the sample and population:
- An allergy institution contacted 2 079 teenagers who are between 13 and 17 years old and live in South Africa and asked whether or not they used prescribed medications for any mental disorders such as depression or anxiety.
 - A farmer wanted to study the weight of his soybean crop. He randomly picked 100 plants and weighted the soybeans on each plant.
 - A quality control manager randomly selects 50 bottles of Coca Cola that were filled on a specific day to assess the calibration of the filling machine.
13. Chemical and manufacturing plants sometimes discharge toxic-waste materials into nearby rivers and streams. These toxins can adversely affect the plants and animals inhabiting the river and river banks. Researchers conducted a study of fish in the rivers in the Gauteng area. A total of 124 fish were captured, and the following variables were measured for each:
- river where each fish was captured
 - species
 - length (cm)
 - weight (g)
 - concentration of toxins.
- Classify each variable as quantitative or qualitative. If quantitative, also indicate whether it is discrete or continuous. Indicate the measurement scale of each category.
14. A *Mail & Guardian* poll of a sample of South Africans revealed that “85% of those surveyed would choose organically grown produce over produce grown using chemical fertilisers, pesticides, and herbicides”. Is the statement an inferential or descriptive statement? Explain your answer.
15. The owner of a large fleet of taxis is busy with his budget for next year’s operations. A major cost is petrol. To estimate the petrol costs, he needs the total distance his taxis will travel next year, the average cost of petrol and the average petrol consumption of his taxis. The first two figures are provided to the owner, but to obtain the last one, he selected 50 of his taxis and measured the consumption of each:
- What is the population of interest?
 - What is the parameter the owner needs?
 - What is the sample?
 - What are the statistics?
-

UNIT 2

Collection of data

This unit deals with how and where to obtain data that can be used to make informed decisions. The quality of the final product depends on the quality of the raw material used. Researchers have adopted the acronym GIGO – garbage in, garbage out.

After completion of this unit you will be able to:

- distinguish between primary and secondary data sources
- examine various sources of primary data
- obtain an appreciation for the art of questionnaire design
- distinguish between probability and non-probability samples
- obtain the knowledge of conducting a sample
- distinguish between different methods of data collection.

To be effective, data needs to meet the following criteria:

- Data must be available when needed.
- Data should be relevant and useful in making decisions.
- Data must reflect the actual situation. Sufficient checks need to be built in to ensure that errors or deliberate malpractices do not go undetected.
- Data must be auditable. This means that the data can be checked for validity and accuracy once it has been processed, so that any error can be identified and traced.
- Data must be complete. This means that enough data is available to enable you to make decisions about the resulting information.
- It is important to identify the relevant population to find your data, because collecting unnecessary data is time-consuming and costly.

2.1 Sources of data: where to get the data

A statistical study may require the collection of new data from scratch, referred to as primary data, or be able to use already existing data, known as secondary data. It is also possible to use a combination of both sources.

Secondary data is already available in processed form, such as a database, or record keeping within your company that has been collected for some other purpose than we intend to use it for. It is usually available at low cost but you need to be sure that you are

not using unsuitable figures just because they are easily available. Secondary data can be obtained internally or externally.

Internal data comes from within the organisation for its own use, for example from accounting records, payrolls, inventories, sales records, etc.

External data are collected from sources outside the organisation, such as trade publications, the consumer price indexes, newspapers, libraries, universities, official statistics supplied by the Department of Statistics and other government departments, the NIELSEN report on shopping behaviour, stock exchange reports, databases of the Department of Statistics, data on the unemployment rate supplied by the Department of Labour, or data on HIV/AIDS provided by the Department of Health or websites on the Internet.

Primary data is information collected by those wishing to collect their own data. The distinguishing feature of this data is that it will be both reliable and relevant to your purpose. As a result, primary data can take a long time to collect and be expensive. Sources of primary data include observation, group discussions and the use of questionnaires under controlled conditions.

2.2 Primary data sources

You can obtain primary data by:

- conducting an investigation or experiment
- observation
- conducting surveys using questions.

2.2.1 Conducting an experiment

In conducting an experiment, you deliberately impose some treatment on individuals or objects in order to observe the responses. The purpose of an experiment is to study whether the treatment causes a change in the response.

Example 2.1

To determine if there is any relationship between the hours of TV viewing and the channel someone views, we selected a random sample of students and told each one of them which TV channel they must watch over the weekend. Each student recorded the number of hours they watched TV.

2.2.2 Observations

In an observational survey, collecting data relies on watching or listening, and then counting or measuring events as they happen without anything being done to the individuals or objects. Draw up an observation sheet and keep count of the observations using a tally table. We keep count by using straight lines for each item we count (||||). The fifth line is a line across the first four lines so that we can count in multiples of five (#####). The variables of interest are not controlled.

Example 2.2

The Metro Police wanted to determine whether motorists using a certain road wear seatbelts. They observed if the driver used a seatbelt as the cars passed by and counted how many wore seatbelts and how many did not.

The number of motorists wearing seatbelts between 7 am and 8 am on 27 February 2008		<i>f</i>
Wearing seatbelts	### ### ### ###	22
Not wearing seatbelts	### ### ###	15
Total number of motorists		37

2.2.3 Survey by means of asking questions

In this type of survey, you need to decide what questions will be asked and how these questions will be put to the people. Design a list of questions, known as a questionnaire, to make sure that the same questions are asked in the same way. Record the answers in written form. These answers form the basis of your statistical analysis.

Some of the most commonly used methods to collect data when conducting a survey using a well-designed questionnaire are:

- the telephone interview
- the mail questionnaire
- the personal interview.

Telephone interviews involve the presentation of the questionnaire by telephone. Telephone surveys are less costly than personal interviews and we can conduct them over wider geographical areas. Another advantage is that you can do more interviews in a given period.

First, select a random sample, then make use of professional telephone operators to dial the numbers of the sample, and asked predetermined questions. People are more open in their opinions as there is no face-to-face contact. One of the major drawbacks is that some people in the sample will not have phones or will not be home when you call them.

In the **mail questionnaire survey**, respondents are asked to complete and return a questionnaire, which they receive in the mail, newspaper, magazine or attached to a product. Use this method to cover a wider geographical area rather than telephone or personal interviews since it is the least expensive. In the same amount of time, a larger sample can be reached. Respondents can remain anonymous if they desire and will therefore be more open and honest in their opinion. Disadvantages of this method include the low response rate, inappropriate answers to questions and illiteracy of some people included in the sample.

During a **personal interview**, the data is obtained verbally and face to face. Interviewers select candidates in a random way from appropriate places such as the university campus or shopping centres. They can select their sample from a certain age group or from people buying groceries. This method is popular with companies conducting market research about certain products. The interviewer must tell the

respondent beforehand how long the interview will take, otherwise the randomly selected respondent will try to avoid the interview. Interviewers must be trained to ask questions and record responses, which make this method more costly and time consuming. An advantage of this method is that you can obtain in-depth responses from respondents, not only by listening to the answer but also by interpreting their body language. The interviewer can clarify difficult questions and show visual displays or products to the respondent to provide better communication and motivation to participate in the survey.

Activity 2.1

Rate the survey methods as either 1 - most appropriate; 2 - less appropriate; 3 - least appropriate, under the following circumstances:

	Telephone interview	Personal interview	Mail questionnaire
Large geographical area			
Small sample			
Difficult questions			
Keeping the cost low			
Body language			
If speed is a factor			
Response rate			
Illiteracy			
Training of interviewers			
Confidentiality			
Market research for a product			

2.3 Questionnaire design

The basis of statistical analysis will be the data obtained in response to questions. You now need to decide on the questions that will be used, and how these questions will be asked.

To be successful, a questionnaire needs both a logical structure and well thought-out questions. The structure of the questionnaire should ensure that there is a flow from question to question. Any radical jumps between topics will tend to disorientate the respondent and will influence the answers given.

A questionnaire can be divided into different parts:

- Administrative part: date, name, address, etc.
- Classification part: race, sex, age, marital status, occupation, etc.
- Subject matter of inquiry (questions)

2.3.1 Objectives of a question

1. To find out if the respondent is aware of the issue, for example "Do you know of any plans to build a highway through the Kruger National Park?"

Yes	No
-----	----

2. To get general feelings on an issue, such as "Do you think the highway should be built?"

In constructing such a question, you can ask the respondent to provide an answer on a rating scale such as:

Strongly agree	Agree	Uncertain	Disagree	Strongly disagree
----------------	-------	-----------	----------	-------------------

3. To indicate feeling you can also use symbols.

Yes 😊	Not sure 😐	No 😞
----------	---------------	---------

4. To get answers on specific parts of the issue, for example "Do you think a highway will affect the local environment?"

Yes	No
-----	----

5. To get reasons for a respondent's view, for example "If against, is it because
 - the highway will spoil the nature
 - the highway will disturb the animals
 - there is an adequate main road already
 - other?"

Alternatively, you can use an open-ended question to get reasons, for example "Why are you against building the highway?"

6. To find how strongly these views are held.

Which of the following will you do to support your view?

- Write to the Director of the National Parks Board.
- Sign a petition.
- Write to a national newspaper.
- Disrupt the construction work.

How important is the conservation of wildlife to you?

Rate your answer: should be done at any cost is 1 and of no importance is 4.

1	2	3	4
---	---	---	---

7. Question wording

When formulating the question, make every effort to ensure that the wording meets the following criteria:

- The question should be clear to the respondent and not open to misinterpretation. Use terms or vocabulary that the respondent understands. If you want to know the respondent's name, specify if it is his or her first name, name and surname, initials and surname, or just a nickname.

- Questions should be short, simple and to the point.
- Do not ask too many questions or questions that are too long, because few respondents will be prepared to spend much time answering the questions.
- Questions should not require any calculations.
- Questions should not lead the respondent. A biased or leading question will bias the answer given – *'bias' means to cause an imbalance*.
- Questions should not be phrased emotively. Place questions that may evoke an emotional response near the end of the questionnaire, since they may influence responses that will follow.
- Questions should not be offensive or embarrass the respondent.
- Wherever possible, a choice of answers should be given (closed questions). When this is not possible, adequate spaces should be given for answers.
- Confidentiality should be assured.

2.3.2 Types of questions we can ask

1. Closed questions give the respondent a series of possible answers from which one must be chosen. This approach makes it easy to record the required information and reduces interviewer's bias.
2. Open-ended questions will allow respondents to give their own opinion in their own words and to express any thoughts that they feel are appropriate to the question. As a result, depending on the nature of the question and the interest of the respondent, the answer may vary a great deal in length and detail.

Activity 2.2

Identify whether the following are open-ended or closed questions:

1. How do you feel about violence in your neighbourhood?
 2. Do you regularly watch soccer on TV? Yes or no
 3. How often do you watch soccer on TV?
 - never
 - sometimes, but not every week
 - one game every week
 - two or more games per week.
 4. What will you do to improve attendance at your school's sporting events?
 5. How reliable is your calculator?
 6. How would you rate the reliability of your calculator?
Superior: Very good: Good: Poor:
-

2.4 Selecting a sample

We can perform any method we choose to collect the data on either a population or a sample.

If we collect data on all the elements of the population, we call it a *census*. The National Census is conducted when each household in South Africa receives a census form to complete, containing information about everybody in that household. If a study

concerns your company, the population will consist of all the employees of your company.

A sample is taken from a sampling frame which is a complete list of people or objects the population consists of.

2.4.1 Advantages of sampling

1. Costs are reduced.
2. Collection time is reduced.
3. Overall accuracy is improved.
4. For several types of populations, sampling is the only method of data collection. For example, infinite populations, or testing procedures that entail the destruction of the item being tested, such as tests determining the life of a light bulb or the length of time a match will burn.

By studying the behaviour of a sample, we can get a good idea of the behaviour of the population from which the sample was drawn. If you summarise and evaluate the sample data you can estimate and draw conclusions about the population parameters based on the sample results or statistics.

Sampling is justified only if the sample is representative of the population and valid inferences about the population can be drawn from the sample. To ensure these two conditions, sampling must be based on two general laws:

1. The Law of Statistical Regularity holds that a reasonable large number of items selected at random from a large group of items will, on the average, have characteristics representative of the population. It is important that the selection of the sample is random so that every item in the population has an equal chance of selection. The size of the sample should be large enough to minimise the influence of abnormal items on the average.
2. The Law of Inertia of Large Numbers holds that large groups of data show more stability than small ones.

2.4.2 Sampling error

We cannot expect that the sample results will be the same as the population results (if known). This difference between the sample statistic and the actual population parameter is known as sampling error. The smaller the sampling error, the more accurate the estimate for the population parameter.

Factors that have an influence on the sampling error:

1. The sample size: the larger the sample, the more similar the sample statistics will be to the population parameter.
2. The amount of variation among the values in the population: suppose you want to investigate the amount of pocket money children receive every month. If these amounts are more or less the same, the variability in the population is small and a small sample will be sufficient. If the amounts differ a lot, the variability is greater and we will need a larger sample.

2.4.3 Sample size (n)

In later study units, a formula will be applied to determine the sample size. For now, we will briefly look at the factors that influence the sample size. The random selection process allows us to be confident that the resulting sample adequately reflects the population, even when the sample consists of only a small fraction of the population.

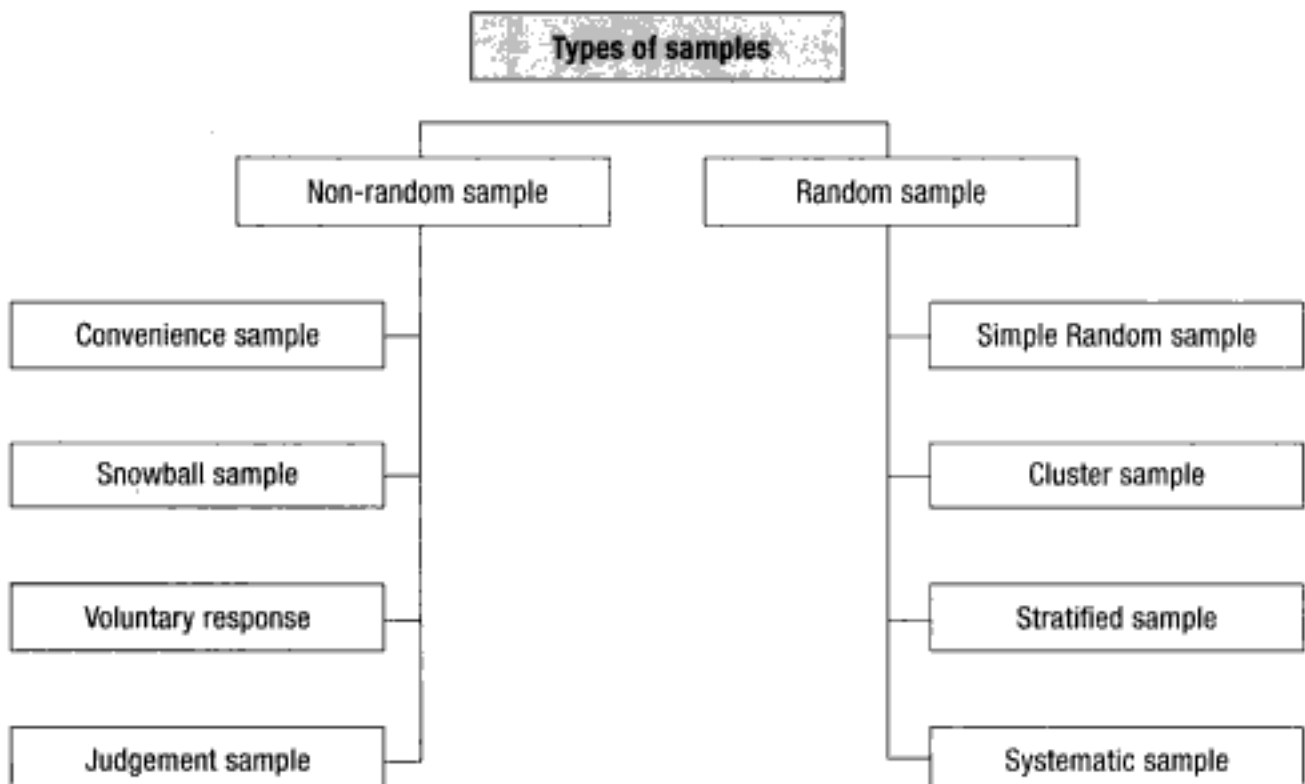
Factors that influence sample size:

1. The size of the population (n). Intuitively we will use a bigger sample for a bigger population, but this factor is less important than the other factors listed.
2. Resources available. This includes time, money and other resources. Less time available will result in a smaller sample.
3. Error that can be tolerated. The bigger the sample you select, the more accurate the conclusions you draw about the population based on the sample will be.
4. Variation in the population. Suppose you wish to study all the students in your university, but all the students feel exactly the same about the issue you are investigating. You want to be very accurate in your conclusion. How many students do you need to talk to? A sample size of one is adequate despite the need to be very accurate and the large size of the population. The more similar the elements of the population are to one another, the less variation there is within the population, so the sample can be smaller.

2.4.4 Sample design

The design of a sample describes the method used to select the sample from the population.

Sampling design can be divided into two broad categories: those where elements are selected by some random method and those where the elements are non-randomly selected.



2.5 Non-random sampling

If the sample items are selected using personal convenience, expert judgement, or any type of conscious researcher selection, the sample selection is not done by chance and is called a non-random sample. Samples like these often produce unrepresentative data and are not desirable for use in inferential statistics. Techniques that follow non-random selection of data include convenience sampling, judgement sampling, voluntary response sampling and snowball sampling.

2.5.1 Convenience sampling

The researcher chooses elements that are readily available, nearby or willing to participate. It is convenient for the researcher to select the first few sample items quickly. When both time and money are limited, convenience samples are widely used. For example:

1. man-in-the-street interviews
2. lunch-hour interviews
3. interviewing close friends or family
4. door-to-door interviews.

2.5.2 Judgement sampling

These samples consist of items deliberately chosen from the population on the basis of the experience and judgement of the researcher. This method usually results in making systematic errors in one direction. These systematic errors lead to what are called biases. For example, four of the most influential economists were asked to estimate next year's rate of inflation.

2.5.3 Voluntary response sampling

These samples consist of people who choose themselves by responding to a broad appeal, such as online polls or newspaper questionnaires. People who take the trouble to respond to an open invitation are usually not representative of any clearly defined population, because only people with strong opinions are most likely to respond.

2.5.4 Snowball sampling

Sample elements are selected based on referral from other survey respondents. The researcher identifies a person who fits the profile wanted for the study. The researcher then asks this person for the names and locations of others who also fit this profile. Through these referrals, sample elements can be identified cheaply and efficiently, which is particularly useful when survey subjects are difficult to locate.

2.6 Random sampling

A random sample is one in which the items chosen are based on chance – the procedure must be such that every element of the population has the same chance (or probability) of being selected into the sample. The four basic random sampling techniques are

simple random sampling, systematic random sampling, stratified sampling and cluster sampling.

2.6.1 Simple random sampling

This technique is the basis for the other random techniques. Each unit of the sampling frame is numbered from 1 to N (where N is the size of the population), or assign any ID number to each element in the population.

Two of the major random techniques are:

1. The '*goldfish bowl*' technique, which is similar to drawing names from a hat. This method works well with a small sample. Place a numbered card for each element in the population in a bowl, mix them thoroughly, and select as many cards as needed in the sample. This method is used often in lottery draws or where the population is small.
2. **Table of random numbers.** Random number tables consist of rows and columns in which the numbers 0–9 appear. A random number generator is a computer program that generates these numbers. Any series of numbers read across or down the table is considered random.

Example 2.3

Assume that you have 100 employees in a company and wish to interview a random sample of 10. Assign every employee a number from 00 to 99. You assign a two-digit number to each element in the population, and then you can use two digits of each number from the random number list. The first step in selecting a sample is to decide where in the random table you should start. Use the random table given below. You can choose to use the first two digits, the middle two digits or the last two digits. You can even choose which columns to use. You can make this decision by using the 'goldfish-bowl' technique or by closing your eyes and pointing to a spot in the table. Suppose you have decided to start in the first column with the first two digits, and the population consists of numbers from 00 to 99. If we reach the bottom of the last column on the right and are still short of our desired number, we can go back to the beginning and start reading the third and fourth digits of each number. According to the table, employee numbers 70, 23, 20, 22, 53, 39, 48, 64, 12 and 45 will be in the sample of ten.

Note that if a number occurs more than once, you skip it. You can't use any population ID twice because there is a unique ID assigned to each element in the population.

Activity 2.3

Each student at the university has a mailbox on campus. The mailboxes are numbered from 0 000 to 9 000. Use the random number table and select 10 mailbox numbers in your sample. Compare your results with some of the results obtained by other students in the class and comment on your findings.

Random number table

7081	8887	2876	1705	4260	5065	5528	8241	5997
2318	0139	6986	4900	2408	2027	1676	4382	3370
2099	3526	7912	3824	5108	1033	7363	0183	8479
2293	4424	9209	5979	5022	4849	1960	1771	7961
5359	3108	7453	9978	3538	8963	9562	5437	6806
3971	9260	0760	1284	1020	0961	2666	0255	5957
4833	6395	4528	0665	5386	3539	5918	9165	2088
6492	9493	1058	9069	7725	0094	9513	2735	2915
1227	1585	3239	0593	4703	4737	5851	2551	2824
4505	9108	0031	9578	0077	9836	5817	3221	1174
9515	4576	4486	8388	1343	4507	0031	2209	1921
9889	6933	2616	3883	9008	3389	3672	6952	5839
5737	6911	3388	3682	7271	1110	7272	5674	1650

2.6.2 Stratified sampling

Identify non-overlapping groups or strata within the population before sample selection takes place. Select a simple random sample from each stratum. Make sure that each of these groups is represented proportionally in the sample. For example, if a researcher needs to estimate the average mass of a large group of people, he or she first divides the group into two strata - male and female - and then selects a proportional simple random sample from each stratum.

2.6.3 Systematic sampling

Select the starting number (a value between 1 and k) at random and each successive number systematically from an orderly list of the population to obtain the sample. Every k^{th} item is selected to produce a sample of size n from a population of size N . The value of k can be determined by the following formula:

$$k = \frac{N}{n}$$

2.6.4 Cluster sampling

Some populations have non-overlapping areas or groups, which within themselves represent all of the views of the general population, for example a town, university or a file of invoices. If this is the case, it will be much more convenient and cost effective to select one or more of these clusters at random and then to select a sample from the clusters, or carry out a census within the selected clusters. Sometimes the clusters are

too large and a second set of clusters is taken from the original chosen clusters. This technique is called multi-stage sampling.

A large geographical area is often divided into more manageable provinces or clusters. Select a few provinces and then select a few towns from each province. Out of each town select a few blocks, and out of each block select individual families at random.

TEST YOURSELF 2

1. "How much do you trust information about health that you find on the internet?" You want to ask a sample of 10 students chosen from your class the question. Describe how you will select your sample using a random method.
2. You want to select a random sample of 25 of the approximately 371 active telephone area codes covering South Africa. Explain the method you will use and select your sample.
3. At a party there are 30 students over the age of 21 and 15 students under age 21. You want to select a representative sample of five to interview about attitudes toward alcohol. Explain your method and select your sample.
4. Based on satellite images, a forest area in KwaZulu Natal is divided into 14 types. The area of each type is divided into large sectors. Chose 18 sectors of each type at random and count the tree species in a 20×25 meter rectangle randomly placed within each sector selected. Explain the method you will use and select the sectors.

Forest type	Total sectors	Sample size
A	36	4
B	72	7
C	31	3
D	42	4

5. You want to choose four addresses at random from a list of 120 addresses. Use the systematic method and describe how you will obtain your sample.
6. The New Firearm Policy Survey asked respondents' opinions about government regulation of firearms. If you are the researcher, and you want to follow the telephone interview method using the multistage cluster sampling method, how will you go about selecting your sample?
7. In the 1940s the public was greatly concerned about polio. In an attempt to prevent this disease, Jonas Salk of the University of Pittsburgh developed a polio vaccine. To test the vaccine, 1 000 000 children received the Salk vaccine and another 1 000 000 a placebo, in this case an injection of salt dissolved in water. Neither the children nor the doctors performing the diagnoses knew which children belonged to which group, but an evaluation centre did. The centre found that the incidence of polio was far lower among children inoculated with the Salk vaccine. From that information, the researchers concluded that the vaccine would be effective in preventing polio for all school children and made it available for general use.

Is this investigation an observational study or a designed experiment? Justify your answer. Is the conclusion of the researchers descriptive or inferential?

8. An inspector of the Department of Food and Drug Administration obtains all vitamin pills produced in an hour at the Herbal Supply Company. She thoroughly mixes them, and then scoops a sample of 10 pills that are to be tested for the exact amount of vitamin content. Does this sampling design result in a random sample? Explain.
-

UNIT 3

Summarising data using tables and graphs

In this unit you begin your study of descriptive statistics. We will look at ways to describe data by summarising and displaying it using tables and graphs so that the salient features of the dataset are more easily understood.

After completion of this unit you will be able to:

- recognise the difference between grouped and ungrouped data
- construct a frequency distribution
- draw graphs based on qualitative and discrete data
- draw graphs based on continuous data
- recognise the usefulness of visual aids in presenting data.

Once data are collected, they must be organised and summarised, because we are concerned with the overall picture rather than treating each observation individually. When data are obtained, the initial result is usually a list of the observations for each variable. This is referred to as raw data. Raw data provide little information. Statistics give us some tools or techniques for turning raw data into useful information. Thanks largely to the advances in personal computing, managers these days have a considerable facility to display information visually and to focus quickly on the key characteristics of a set of data.

The major principles of visual data presentation:

1. Approved methods should be used to confirm the reliability and truthfulness of the presentation, so that the reader can interpret it correctly.
2. The presentation must be as simple as possible so that the reader may understand it quickly and easily.
3. The presentation must interest the reader and contain worthwhile information.
4. An explanation should be included in the text in order to help the reader understand the purpose of the presentation.

The stages to follow in summarising data in tables and graphs:

1. Order the data into a logical sequence.
2. Summarise it by grouping and arranging data in the form of a table known as a frequency distribution.
3. Present it in an attractive way, using graphs or diagrams. The choice of presentation depends on the type of data, the complexity of the data and the requirements of the user.

A **graph** shows the relationship between two variables, one will be the x -variable on the horizontal axis and the other the y -variable on the vertical axis. A graph does not replace a table, but complements it by showing the data's general structure more clearly and revealing trends or relationships that might be overlooked in a table. It is more likely to receive the attention of the casual observer. The type of graph depends on the type of data, the complexity of the data and the requirements of the user.

3.1 Summarising qualitative data in tables and graphs

3.1.1 Frequency distributions

A **frequency distribution** is a table that records each category, value, or interval of values that a variable might have and the number of times (**frequency**) that each one occurs in the data set.

A **frequency distribution** for qualitative data is a table that displays the possible categories along with the corresponding frequencies. The frequency for a particular category is the number of times the category appears in the data set.

The **relative frequency** for a particular category is the fraction, proportion or percentage of the observations within a category. If the table includes relative frequencies, it is referred to as a **relative frequency distribution**.

You can use a fraction, decimal or percentage to express the relative frequency, but percentages are easiest for most people to understand.

$$\text{relative frequency} = \frac{\text{frequency}}{\text{sum of all frequencies}}$$

Steps

1. Construct a column in which each row lists one of the categories for the variable of interest.
2. Construct a second column to list the corresponding number of times that the category occurs.
3. Add up the frequency column to make sure that the total is the same as the number of observations.
4. The order for the categories in the frequency table is not important, unless there is a logical order in the given data set.
5. Interpret the table results.

Example 3.1

Toni's Supermarket has received many complaints about the condition of cardboard milk boxes. Customers are refusing to buy boxes that are dented and crushed, because they don't know whether the contents are still intact. Since the store manager is fairly sure that the damage is not occurring when the boxes are put on the shelves, he goes to the warehouse to check the cases as they arrive from the distributor. He takes a random

sample of milk in boxes as they arrive and examines them for various defects. The sample provides the following data:

Unsealed	No defect	Dented	Crushed	No defect	Crushed	Unsealed
No defect	Dented	No defect	No defect	Dented	No defect	No defect
Dented	No defect	No defect	Crushed	Crushed	Crushed	No defect
Crushed	No defect	No defect	Dented	No defect	Crushed	No defect

1. Create a frequency table for the data and determine if his concerns are justified.
2. Change the frequency distribution to a relative (%) frequency distribution.

Category	Frequency (f)	% f
Unsealed	2	7
Crushed	7	25
Dented	5	18
No defect	14	50
Total	28	100

Conclusions:

1. The table shows that half of the boxes that arrived were damaged which is definitely a matter for concern. Only two of the boxes were unsealed and can be considered as unsafe for use.
2. 50% of the boxes are damaged with half of the damaged boxes crushed.

Activity 3.1

A bio-kinetics instructor wants to study the different types of rehabilitation required by her patients. She selects a simple random sample of her patients and records the body part requiring rehabilitation. The following results are obtained:

Hand	Back	Ankle	Shoulder	Back	Back
Back	Shoulder	Back	Wrist	Knee	Knee
Neck	Ankle	Hip	Knee	Back	Neck
Wrist	Shoulder	Back	Back	Back	Shoulder
Knee	Back	Back	Knee	Hand	Wrist

Construct a frequency distribution and a relative frequency distribution to describe the data. Give a short interpretation of your results.

3.1.2 Cross-tabulation

Data resulting from observations made on **two** different related categorical variables (bivariate) can be summarised using a table, known as a two-way frequency table or contingency table.

The word contingency is used because the table is used to determine if there is an association between the variables.

Steps

1. This table displays the one variable (x) in the rows and the other variable (y) in the columns.
2. Each row and column combination in the table is called a cell.
3. The number of times each (x, y) combination occurs in the data set is recorded and these numbers are entered in the corresponding cells of the table. These are known as the **observed cell counts**.
4. Add the observed cell counts in each row and also in each column of the table to obtain the **marginal totals**.
5. The **grand total** is the total of all the observed cell counts in the table. All the row marginal totals will add to the grand total. All the column marginal totals will also add to the grand total.
6. We use a contingency table if we want to compare two different populations on the basis of a single categorical variable, or when two categorical variables are observed in a single sample. For example, data could be collected at a university to compare students, staff and management on the basis of their transportation to campus (taxi, bus, car, train, motorcycle, bicycle or on foot). This will result in a (3×7) two-way frequency table with row categories of student, staff and management, and column categories corresponding to the seven possible modes of transportation. The observed cell counts could then be used to gain insight into differences and similarities among the three groups with respect to the means of transportation.

Activity 3.2

People believe that organic foods are pesticide-free and thus healthier than conventional grown fruit and vegetables. The organic fruit and vegetable market is however, very small and therefore very expensive. An investigation is carried out on a sample of 26 000 food items as part of the regulatory monitoring of foods for pesticides residues before being distributed for sale. The following table displays the frequencies of foods for all possible category combinations of the two variables: food type and pesticide status.

	Pesticides		
Food type	Present	Not present	Total
Organic	28	99	127
Conventional	19 085	6 788	25 873
Total	19 113	6 887	26 000

Briefly comment on these results.

Activity 3.3

One hundred students majoring in Sciences were classified according to gender and year of study. Ten were 1st-year women, 20 were senior women, 40 were 1st-year men and 30 were senior men. Arrange the data in a bivariate table.

3.1.3 Bar graph for a single data set

Bar graphs are a quick and easy way of showing variation in or between variables.

One of the axes is used to represent the categories in the frequency table and the other axis is used to represent the frequencies or relative frequencies. Single bars representing each variable are drawn either vertically or horizontally. Assignment of axes is a matter of preference, but for the purpose of uniformity, we will use the x -axis to represent the categories.

Steps

1. Communicate only a single idea or variable.
2. Draw a pair of axes, x and y .
3. Label the axes and give the graph a title.
4. At evenly spaced intervals on the x -axis put tick marks and label them with the categories from the frequency table.
5. Scale the y -axis so that it can accommodate the category with the highest frequency or relative frequency. Whenever you use a change of scale in a graph, indicate it by using a squiggle ↯ or //.
6. At each category on the x -axis draw a bar with its length equal to the frequency or relative frequency for the variable it represents.
 - The bars must all have the same width.
 - Make the bars reasonably wide so that they can be clearly seen.
 - The gaps between the bars must have the same width - the bars should not touch each other.
7. This type of graph not only illustrates a general trend, but also allows a quick and accurate comparison of one period with another or illustrates a situation at one particular time.

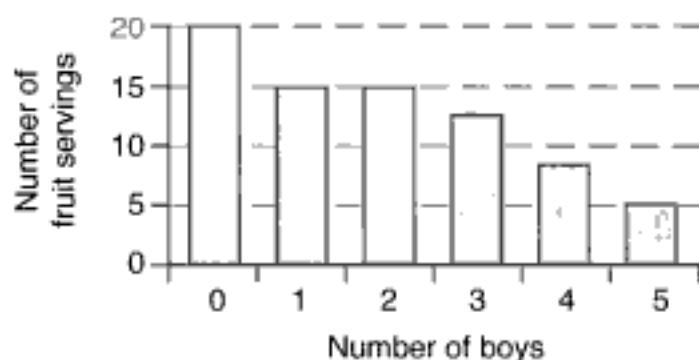
8. If you arrange the bars in a decreasing order, the graph is called a Pareto Chart. By arranging bars in order of frequencies, the attention is drawn to the more important categories.

Example 3.2

We all know that fruit is good for us and that we don't eat enough. In a recent study done among a random sample of 75 teenager boys, the following information was collected:

Fruit servings per day	Number of boys	% of boys
0	20	27
1	15	20
2	15	20
3	12	16
4	8	11
5	5	6

1. Display these data in a bar graph.
2. If the number of recommended servings per day is at least three, what percentage of the boys ate fewer than 3 servings per day?



Conclusion: 67% of the boys ate less than the recommended number of servings.

Activity 3.4

Draw a simple bar chart showing the ages of employees and draw conclusions from your results.

Age	Number of employees
20	11
21	4
22	8
23	6
24	5

3.1.4 Comparative bar graphs

To compare two or more data sets, bars are grouped together in each category (multiple bar graphs) or stacked for each category. Use the relative frequency rather than the frequency on the vertical axis, to enable you to make meaningful comparisons even if the sample sizes are not the same. The use of a key will help distinguish between the categories

Multiple bar graph

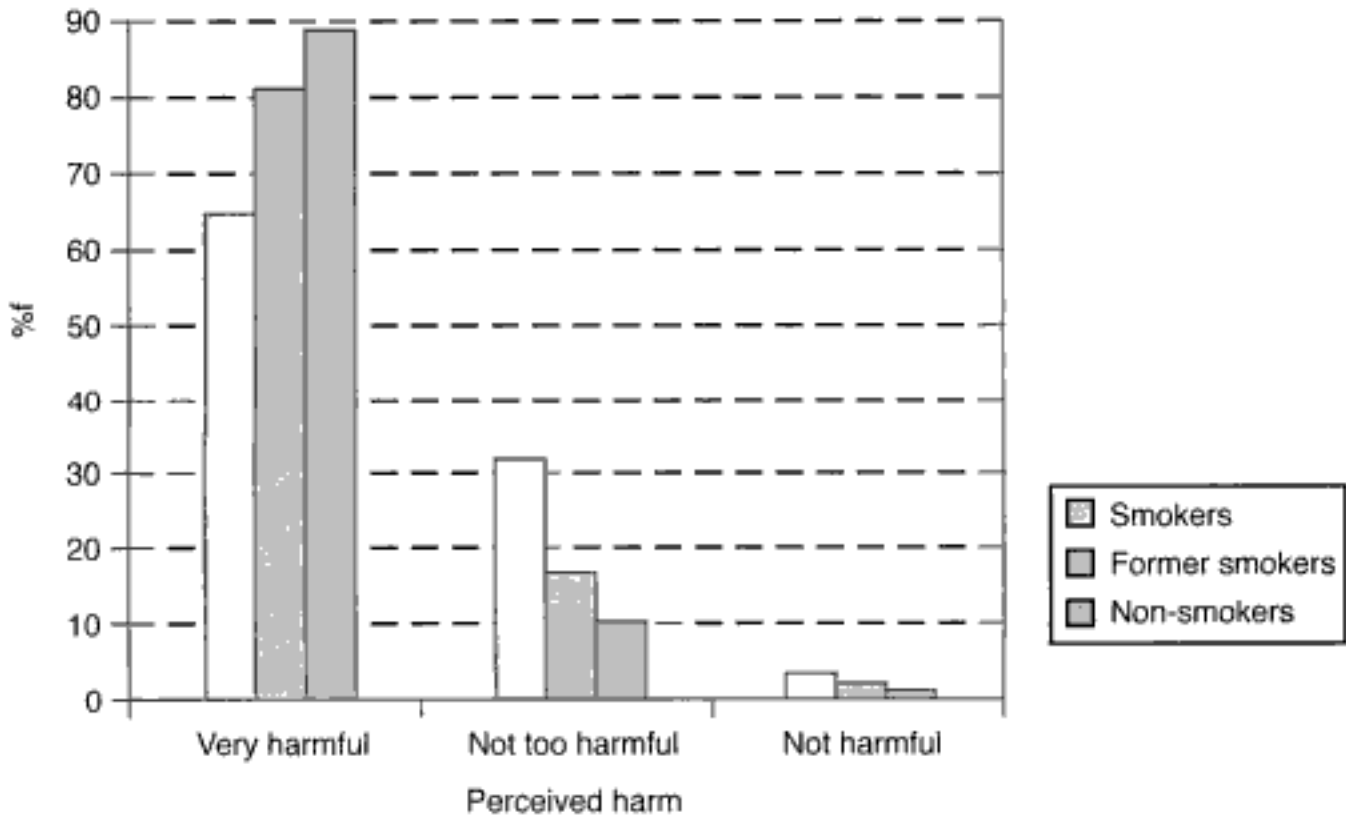
Steps

1. Draw a pair of axes, x and y .
2. Label the axes and give the graph a title.
3. At evenly spaced intervals on the x -axis put tick marks and label them with the categories from the frequency table.
4. Determine the relative frequency of each category if needed.
5. Scale the y -axis so that it can accommodate the category with the highest frequency or relative frequency.
6. At each category on the x -axis, group the bars for the different data sets together and draw rectangles with heights equal to the relative frequency for the data set it represents.
7. Use a key or label to distinguish between the different data sets.
8. Interpret your graph.

Example 3.3

The contingency table below summarises the responses of three different groups to their perceived risk of smoking. Portray the data using a multiple bar graph to determine if smokers and non-smokers perceive the risks of smoking differently.

Risk of smoking	Smokers		Former smokers		Non-smokers	
	<i>f</i>	% <i>f</i>	<i>f</i>	% <i>f</i>	<i>f</i>	% <i>f</i>
Very harmful	145	65	204	81	432	89
Not too harmful	72	32	42	17	50	10
Not harmful at all	7	3	5	2	5	1
Total	224	100	251	100	487	100



The graph shows that the proportion of non-smokers who believe that smoking is very harmful is larger than the proportion of smokers who believe that smoking is very harmful. In other words, smokers are less likely to believe that smoking is very harmful than non-smokers.

Activity 3.5

Draw a multiple bar chart showing the ages of male and female employees and comment on your result.

Age	Male	Female
20	3	8
21	1	3
22	4	4
23	2	4
24	1	4

Segmented or stacked bar chart

This bar chart is particularly useful if you want to emphasise the relative proportions of each component that makes up the category.

Steps

1. Draw a single bar for each category, with the height of the bar representing the total of each category.
2. Subdivide each bar to show the components that make up each category.
3. Identify the components involved by colouring or fill effects, accompanied by an explanatory key to show what each colour or fill effect represents.
4. Interpret your results.
5. If the components are converted to percentages of the total of each category, the bars are divided in proportion to these percentages. The scale is a percentage scale and the height of each bar is then 100%. This is known as a percentage component bar graph.

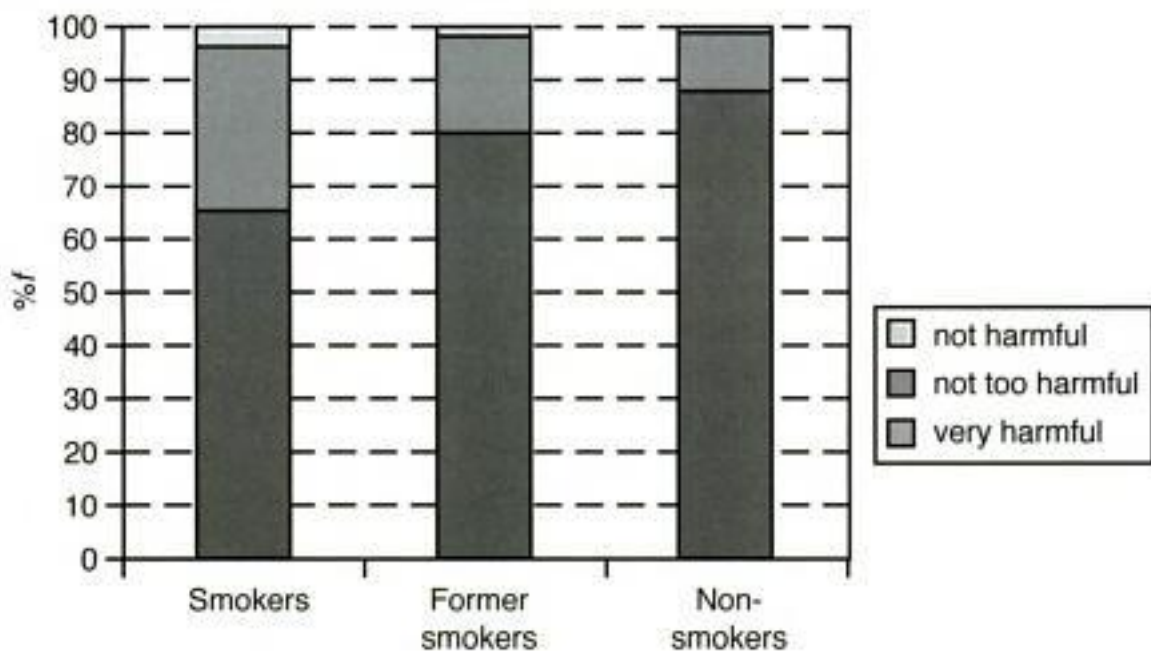
Example 3.4

The contingency table below summarises the responses of three different groups to their perceived risk of smoking. Portray the data using a percentage component bar graph to determine if smokers and non-smokers perceive the risks of smoking differently.

Risk of smoking	Smokers		Former smokers		Non-smokers	
	<i>f</i>	% <i>f</i>	<i>f</i>	% <i>f</i>	<i>f</i>	% <i>f</i>
Very harmful	145	65	204	81	432	89
Not too harmful	72	32	42	17	50	10
Not harmful at all	7	3	5	2	5	1
Total	224	100	251	100	487	100

In comparing the three columns, we see that the proportion of smokers who believe that smoking is very harmful is much larger than the proportion who believes that it is not very harmful or not harmful at all. There is a sharp increase in the non-smokers'

belief that smoking is very harmful, with a decrease in the proportions who believe that it is not too harmful or not harmful at all.



Activity 3.6

Draw a stacked bar chart showing the ages of male and female employees and comment on your result.

Age	Male	Female
20	3	8
21	1	3
22	4	4
23	2	4
24	1	4

3.1.5 Pie chart

A pie chart represents the data set in the form of a circle divided into 'slices' representing the possible categories. This allows a quick overall view of the relative sizes of the categories, but offers little potential for comparison.

Pie charts are most effective for relatively simple representations and summarising data sets when there are not too many categories.

Steps

1. Draw a circle to represent the entire data set.
2. Keep the categories to 10 or fewer.
3. For each category calculate the 'slice' size.

- A circle has 360° and 'slice' sizes are calculated as a proportion of 360° . 'Slices' are drawn by making use of a protractor.
- Put any labelling outside the circle.
- Look for categories that form large and small proportions of the data set when interpreting the chart.

Example 3.5

A random sample of 2 000 shoppers was asked why they were visiting a shopping centre on a specific day.

	Number of shoppers	%f	$^\circ$
Groceries	790	0.395	142
Clothing	570	0.285	103
DIY	580	0.29	104
Other	60	0.03	11
Total	2 000	1	360

A pie chart showing the main purpose of shopping



The majority of shoppers on that specific day wanted to buy groceries. Equal proportions wanted to buy clothing or DIY items, and only few people were there for other purposes.

Activity 3.7

Here is how you might divide up your day:

Travelling	Working	Eating	Sleeping	Other	Social life
10%	30%	10%	28%	7%	15%

Draw a pie chart to portray the data and comment on the results.

3.1.6 Pictograms

Pictograms are small symbols or simplified pictures that represent data.

Steps


1. Give the pictogram a title.
2. Choose a simple symbol or picture that is easy to draw.
3. The quantity that each symbol represents should be given.
4. It is important that the symbols are all the same size. It is possible to use half a picture to represent half the quantity.
5. Draw the symbols neatly and professionally.

The number of telephone calls received

(1 unit = 100 calls)

January = 

February = 

March = 

3.2 Summarising quantitative data in tables

3.2.1 The ordered array of data

If there are not too many observations we can use the collected data in its raw form, known as ungrouped data.

A first step in organising ungrouped data is to arrange the data in an array – that means to sort the data in numerical order from small to big. By looking at an ordered array, you can get a feel for the dimension of the data. Data must be in order for a variety of statistical procedures, such as finding the median, percentiles or quartiles.

Example 3.6

Arrange the following data in an array:

4 80 50 10 5

Array: 4 5 10 50 80

Activity 3.8

Arrange the following numbers in an array:

67 23 56 45 56 41 34 33 0 18 23

3.2.2 Dot plot

This method can be used for relatively small data sets (usually not more than 20 observations) and portrays individual observations.

Steps

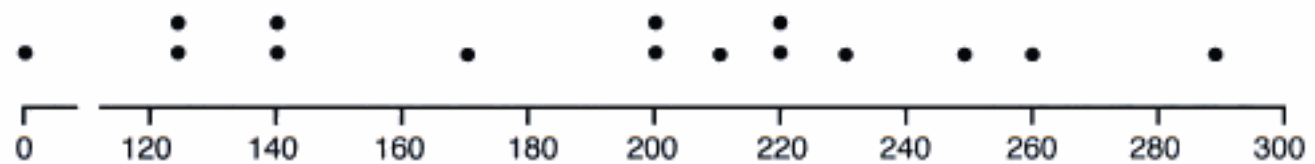
1. Construct a single horizontal axis and label it with the name of the variable.
2. Mark the axis with an appropriate measurement scale to fit the smallest as well as the largest value in the data set.
3. For each observation, place a dot above its value on the number line.
4. If there are two or more observations with the same value, stack the dots vertically.
5. The number of dots above a value on the number line represents the frequency of occurrence of that value.

Example 3.7

Obesity, high blood pressure, high cholesterol and heart disease are partially caused by a poor diet. Nutritional labels on packaged foods give us information about the amount of fat, cholesterol, sodium, vitamins and carbohydrates contained in a serving of the food. The purpose of this study is to investigate how much sugar and how much sodium (the main ingredient of salt) is in breakfast cereals. The following table lists 15 popular cereals and the amounts of sodium and sugar contained in a single serving of 180 ml.

Cereal	Sodium (mg)	Sugar (g)	Cereal	Sodium (mg)	Sugar (g)
A	290	2	I	250	10
B	200	3	J	125	14
C	230	3	K	220	3
D	125	13	L	0	7
E	260	5	M	220	12
F	200	11	N	170	3
G	210	12	O	140	10
H	140	10			

Construct a dot plot for the sodium values of the breakfast cereals.



What does the dot plot tell us about the data?

The dot plot gives us an overview of all the data. We see that the sodium values fall between 0 and 290 mg, with most cereals falling between 125 and 250mg.

Activity 3.9

1. Construct a dot plot for the sugar values of the breakfast cereals.
2. What does the dot plot tell us about the data?

3.2.3 Stem-and-leaf plot

This graph portrays the individual observations and provides a fast procedure for arranging data in order and showing the shape simultaneously. Use it for data sets with a small to moderate number of observations.

An advantage of this method is that all the information in the original data list is shown and if necessary, we could reconstruct the original list of values.

The stem-and-leaf plot represents data by separating each value into two parts: the stem and the leaf.

The stems are the leading digit or digits and are displayed in a vertical position on the left hand side of a vertical line. Usually the stem consists of all the digits except for the final one, which is the leaf or trailing digit. To display the value 76 into this format, the 7 will be the stem and the 6 will be the leaf.

Stem	Leaf
7	6

Units of measure: Stem: tens
Leaf: ones

Steps

1. Select one or more leading digits for the stem values. You can choose the digits to serve as the stem, but keep them constant for all the stems.
2. Find the smallest number and the largest number in the distribution of numbers. These will give the first stem and the last stem.
3. List all possible stems in increasing order, to the left of the line.
4. The trailing digit(s) become the leaves.
5. Record the leaf for every observation beside the corresponding stem value.
6. Place the leaves with the same stem on the same row as the stem.
7. Arrange the leaves in each row from lowest to highest to form a stem-and-leaf plot.
8. Use a label to indicate the units for stems and leaves in the display.
9. Count the number of leaves per row and enter the answer in a column next to the display. That is the frequency of each row.
10. The display conveys information about:
 - a representative or typical value in the data set
 - the extent to which the data values are spread out
 - the presence of any gaps in the data
 - the extent of symmetry in the distribution of values
 - the number and location of peaks
 - the presence of unusual values (outliers) in the data set.

11. When the stem has many leaves, it does not clearly portray where the data falls. In this case it is useful to split each stem in two: putting leaves from 0 to 4 on the first stem and from 5 to 9 on the second stem.
12. To make a stem-and-leaf plot more compact, we can remove the last digit. For example, 0.311, 370 and 125 will become 0.31, 37 and 12. Just remember to indicate the correct unit for the leaves, for instance in the case of 125, if the 5 falls away the stem will be 1 with unit hundred and the leaf will be 2 with unit ten.

Example 3.8

Construct a stem-and-leaf plot for the test marks obtained by a sample of 20 students.

78 82 96 74 52 68 82 78 74 76
88 62 66 76 76 84 95 91 58 86

1. The smallest number is 52 and the largest number is 96. Use the first digit (the tens) in each number as the stem and the last digit (the units) as the leaf.

Stem	Leaf
5	
6	
7	
8	
9	

2. Place each leaf on its stem by placing the trailing digit of each data value on the right side of the vertical line opposite its corresponding leading digit (stem). The first value is 78 with 7 the stem and 8 the leaf. Thus, we place 8 opposite the stem 7.

Stem	Leaf
5	25
6	826
7	8484666
8	22684
9	651

3. Order the trailing digits in each row from lowest to highest to form a stem-and-leaf plot.

Stem	Leaf
5	25
6	268
7	4466688
8	22468
9	156

4. To focus on the shape indicated by the stem-and-leaf plot, use a rectangle to contain the leaves of each stem and rotate the page onto its side. A picture similar to a histogram is seen.

Stem	Leaf
5	25
6	268
7	4466688
8	22468
9	156

The general shape is almost symmetrical around the seventies and the majority of the students obtained marks of 70 and above.

5. Count the number of leaves per row and enter the answer in a column next to the display. That is the frequency of each row.

Stem	Leaf	Frequency
5	25	2
6	268	3
7	4466688	7
8	22468	5
9	156	3
		20

Activity 3.10

Given is an array of the daily litres of used sunflower oil bought by a bio-diesel plant. Construct a stem-and-leaf plot for the data.

58 63 69 69 70 71 71 72 72 72
 73 73 74 75 77 79 80 82 84 84
 85 88 91 91 91 94 96 97 99 100

3.2.4 Frequency distribution tables

The frequency table condenses the raw data into a more manageable form that will increase our ability to detect pattern and meaning. This is done by keeping count of how many times a particular value occurs. A *frequency* is the number of times a value occurs.

Ungrouped frequency distribution

To demonstrate the concept of a frequency distribution we will use a set of quantitative data and group it into an ungrouped frequency distribution - 'ungrouped' because each x -value in the distribution stands alone.

Example 3.9

15 6 14 15 4 15 17 6 18 15

An array of the x -values and the number of times each one occurs (f):

Value (x)	Frequency (f)
4	1
6	2
14	1
15	4
17	1
18	1

The value 15 occurred four times, therefore it has a frequency of four.

Activity 3.11

1. From example 3.9 above:
 - a) Which values occurred only once?
 - b) The value that occurs the most is ...
 - c) How many values are in the distribution? Count the number of values in the given data set and compare it with the total of all the frequencies.
2. Form an ungrouped frequency distribution of the following data and comment on the frequency of each value:

1 2 1 4 0 2 0 1 4 1 6

Grouped frequency distribution

If the number and range of observed values is relatively large you will have a fairly lengthy list of data, which is not easy to interpret. It is then necessary to summarise the data in a grouped frequency distribution by grouping adjacent x -values into intervals, known as classes. In summarising the values like this we lose the detail of individual values, but it makes the data much easier to read and understand.

A frequency distribution is a summary of numerical data obtained by grouping it into several non-overlapping class intervals showing the number of observations (frequency) in each interval.

Data organised into a frequency distribution using class intervals are called **grouped data**.

Although there are no absolute rules for constructing a frequency table, you can apply some guidelines to help you. Frequency distributions can vary in final shape and design because they are constructed to individual researchers' taste.

Construction of a frequency distribution**Steps**

1. Determine the range of the given ungrouped or 'raw' data. The range (R) is the difference between the largest and smallest values in the data set.
2. Determine the number of class intervals (K). Generally frequency tables should contain between five and 20 classes. As a guideline, the number of classes (K) should be approximately equal to the square root of the sample size, n .

$$K = \sqrt{\text{number of observations}}$$

Round the answer up to the next whole number.

3. Determine the width (c) of the class interval, which is the range divided by the number of classes.

$$c = \frac{R}{K}$$

This answer should be rounded to a whole number or to the same number of decimals as the raw data.

4. Test: The number of intervals multiplied by the width must always be larger than the range. ($K \times c > R$)
5. Choose the lower and upper class boundaries of each interval to indicate the smallest and largest data values that will fall into each class. The classes must span the entire data set and must not overlap.
Begin by choosing a number for the lower boundary of the first class. Choose either the lowest data value or a convenient value that is a little smaller. Add the class width (c) to this value to get the second lower class boundary. Add the class width to the second lower class boundary to get the third, and so on. List the lower class boundaries in a vertical column. The upper class boundary of the first interval is the same as the lower class boundary of the second interval. The last class ends at a value more than the highest number in the range.
6. Sort the raw data into the classes by making use of the tally method. The tally method is a method of counting data that falls into each interval. Examine each data value and determine which class contains the data value. Make a tally mark or vertical stroke beside that class. For ease of counting, each fifth tally mark of a class is placed across the prior four marks (≡ rather than ||||). Observations that fall exactly on the lower class boundary stay in that interval; observations that fall exactly on the upper class boundary go into the next higher class interval.
A class contains all observations from the lower boundary of the class up to but not including the upper boundary.
7. Count the number of tallies (observations) in each class to obtain the frequency (f) for each class.
8. The sum of the frequencies for all class intervals must equal the number of original data values.
9. It is possible to come to some conclusions like: in which class do you find the majority of the values or the least number of values?

Notes

1. The number of classes should be small enough to provide an effective summary but large enough to display the relevant characteristics of the data.
2. Class boundaries must be selected in such a way that the smallest value is included in the first interval and the largest value in the last interval.
3. Avoid overlapping of intervals so that an observation falls in one class only.
4. The width of all classes should be equal.
5. Open-ended class intervals should be avoided although they may be useful when a few values are extremely large or small in comparison with the rest of the values.
6. Class intervals with a frequency of zero should be avoided.

Example 3.10

Research by the Food and Biomedical Administration shows that acrylamide (a possible cancer-causing substance) forms in high-carbohydrate foods cooked at high temperatures and that acrylamide levels can vary widely even within the same brand of food. The researchers analysed Big Mac's French fries sampled from different franchises and found the following acrylamide levels:

366 155 326 187 245 270 319 223 212 190
 193 247 255 235 300 311 180 333 289 245
 328 201 260 259 263 313 151 322 270 299

Construct a frequency distribution for the acrylamide levels.

Range: $366 - 151 = 215$

Number of intervals: $K = \sqrt{30} = 5.47 \approx 6$

Width of interval: $c = 215 \div 6 = 36.83 \approx 36$

Class intervals	Tally	Frequency (f)
151 – <187		3
187 – <223		5
223 – <259		6
259 – <295		6
295 – <331		8
331 – <367		2
	Total f	30

The sample from 30 franchises has been counted into six classes, with a width of 36 each. For example, 151 up to just under 187 is the first class interval, the two numbers 151 and 187 are the class boundaries and three (the number of franchises) is the frequency of that class. This means that in three of the franchises the acrylamide levels in the French fries were between 151 and just under 187.

Note: The Greek capital letter, sigma (Σ), stands for 'sum the appropriate values.'

Thus we write $1 + 2 + 3 + 4 + \dots + n$ as $\sum_{i=1}^n x_i$

This means the sum of all the x values from 1 through n . This index system must be used whenever only part of the available information is to be used. In statistics, however, we will usually use all the available information and the notation will be adjusted by doing away with the index system

$\sum_{i=1}^n x_i = \Sigma x$. Total frequency can therefore be written as Σf .

Activity 3.12

Look again at the data in example 3.10:

1. The number of franchises with acrylamide levels in their French fries between 259 and 295 is ...
2. The frequency for the class with acrylamide levels between 187 and 223 is ...
3. The upper boundary of the first class is ...
4. The lower boundary of the third class is ...
5. The total number of observations in the data set is ...

Activity 3.13

A study was recently carried out to determine the amount of time that non-secretarial office staffs spend using computer terminals. The study involved 50 staff and the times spent using computers, in hours per week, were as follows:

1.2	4.8	10.3	7.0	13.1	16.0	12.7	0.5	5.1	2.2
8.2	0.7	9.0	7.8	2.2	1.8	5.2	14.1	5.5	13.6
12.2	12.5	12.8	13.5	2.5	5.0	15.5	2.5	3.9	6.5
4.2	8.8	7.5	14.4	10.8	16.5	2.8	9.5	17.0	10.5
12.5	10.5	16.0	14.9	0.3	11.6	12.8	17.7	18.0	22.0

Construct a frequency distribution for the data.

Relative frequency distribution

When the proportion of observations in each class interval instead of the actual number of observations is recorded, the distribution is known as a **relative frequency distribution**. Relative frequency distributions are useful for comparing two data sets, especially when the sample sizes or measurement scales differ substantially. A relative frequency of a class is the observed frequency of the class divided by the total number of observations in the data set. If the percentage is required, multiply the result by 100.

Class midpoint or class mark

The class mark or **midpoint** (x) divides a class interval into two equal parts and is obtained by adding the upper and lower boundaries of each class interval and dividing the result by two. This middle value represents the class interval in calculations.

Cumulative frequency distribution

Knowledge of the number of observations that lie below or above a certain value is often desired. A **cumulative 'less than' frequency** for a class is the sum of the frequencies for that class and all previous classes. We read it as the the total of all the frequencies less than the upper boundary of each interval. A **cumulative relative frequency**

distribution is a ratio calculated by dividing a cumulative frequency of a class by the total number of observations in the data set.

Example 3.11

A frequency table showing the acrylamide levels in the French fries from a sample of Big Mac's outlets.

Class Intervals	Frequency (f)	%f	x	cum < f	cum < %f
151 – <187	3	10	169	3	10
187 – <223	5	17	205	8	27
223 – <259	6	20	241	14	47
259 – <295	6	20	277	20	67
295 – <331	8	27	313	28	93
331 – <367	2	7	349	30	100
	30	100			

Interpreting interval 2: 17% of the outlets have acrylamide levels in the French fries of between 187 and 223. Eight of the outlets have acrylamide levels of less than 223. 27% have acrylamide levels of less than 223.

Activity 3.14

Use your frequency table from activity 3.13 and construct a relative frequency distribution, a cumulative frequency distribution, a relative cumulative frequency distribution and the class midpoints.

3.3 Summarising quantitative data using graphs

3.3.1 The histogram and relative histogram

A histogram is a continuous series of rectangles of equal width but different heights drawn to display the class frequencies.

Steps

1. Mark the class boundaries on the x -axis. The class intervals are equal in width; therefore the points must be equidistant from one another.
2. Use either frequency or % f on the y -axis. A proper scale showing the true zero must be used on the y -axis in order not to misrepresent the character of the data.
3. Whenever the zero point on the horizontal axis is not in its usual position at the intersection of the horizontal and vertical axis, the symbol // or some similar symbol should be used to indicate that.

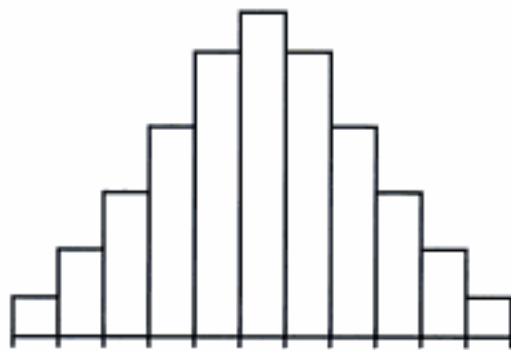
4. Draw a rectangle for each class directly above the corresponding interval. The height of each rectangle is the frequency (or relative frequency) of the corresponding class.
5. There are no gaps between the bars of the histogram.

To interpret the histogram you must look for:

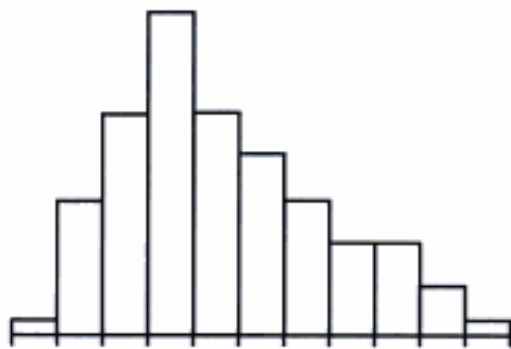
- the overall pattern and obvious deviations from this pattern (the overall pattern can be described by its shape, centre and spread)
- the location and number of peaks
- the presence of gaps and outliers.

Possible shapes of the histogram:

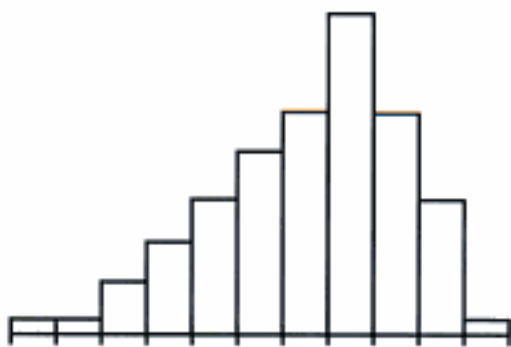
A distribution is **symmetric** if the right hand side is a mirror image of the left hand side:



A distribution is **skewed to the right** if the 'tail' (larger values) extends much farther out to the right:



A distribution is **skewed to the left** if the 'tail' (smaller values) extends much farther out to the left:



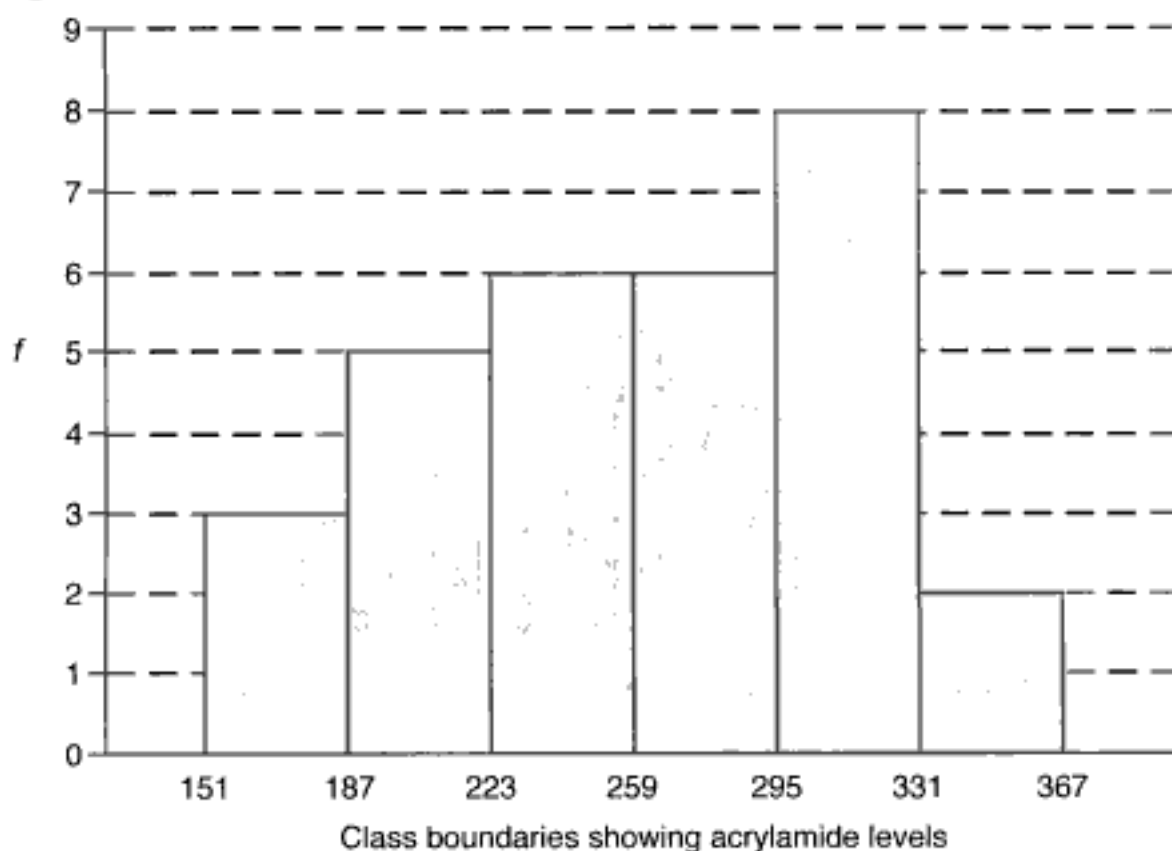
A distribution is **uniform** if the frequency of each class is the same and the bars of the histogram have the same length.

Example 3.12

Draw a histogram showing the acrylamide levels in the French fries from a sample of Big Mac's outlets.

Acrylamide levels	Frequency (f)	$\%f$	x	$cum < f$	$cum < \%f$
151 - <187	3	10	169	3	10
187 - <223	5	17	205	8	27
223 - <259	6	20	241	14	47
259 - <295	6	20	277	20	67
295 - <331	8	27	313	28	93
331 - <367	2	7	349	30	100
	30	100			

Histogram:

**Activity 3.15**

Use your frequency table from activity 3.13 to construct a histogram.

Note: For an ungrouped frequency distribution, the data are grouped into classes based on a single value. To draw the histogram we place the middle of each histogram bar over the single value represented by the class.

3.3.2 Polygon and relative polygon

The polygon is a line graph that can also be used to portray the shape of the distribution.

Steps

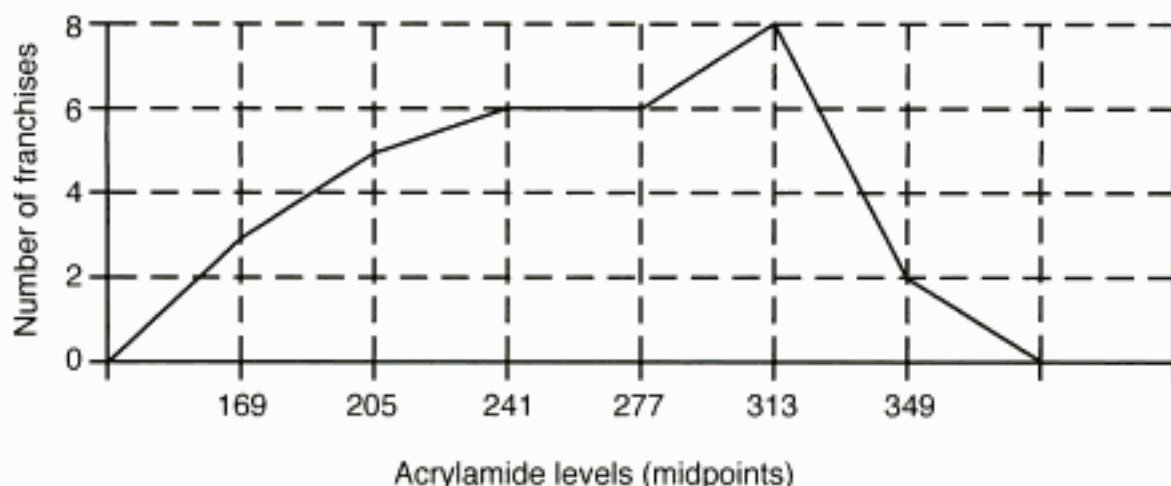
1. The frequency distribution must show the class midpoint (x) of each class.
2. Mark the class midpoints on the x -axis.
3. Mark the frequencies on the y -axis using a proper scale and preferably starting at the zero point. The scale must include values large enough to include the largest frequency.
4. Plot each midpoint together with its corresponding frequency.
5. Connect the successive dots with a straight line to form the polygon.
6. Frequency polygons begin and end on the horizontal axis with a frequency of zero. On the left end plot a point with a frequency of zero, one class width to the left of the first midpoint. On the right end plot a point with a frequency of zero, one class width to the right of the last midpoint.

A polygon that uses the relative frequencies of the intervals rather than the actual number of points is called a relative polygon. It has the same shape as the frequency polygon, but uses a percentage scale on the y -axis.

Example 3.13

Below is a frequency table showing the acrylamide levels in the French fries from a sample of Big Mac's outlets. Draw the polygon for this distribution.

Intervals	Frequency (f)	x	$cum < f$
151 – <187	3	169	3
187 – <223	5	205	8
223 – <259	6	241	14
259 – <295	6	277	20
295 – <331	8	313	28
331 – <367	2	349	30
	30		

Polygon:**Activity 3.16**

Use your frequency table from activity 3.13 to construct a polygon.

3.3.3 Ogive (cumulative curve) and relative ogive

An ogive is a smooth curve that can be used to estimate graphically the number of observations below or above a set level. Therefore an ogive requires cumulative class frequencies.

Steps

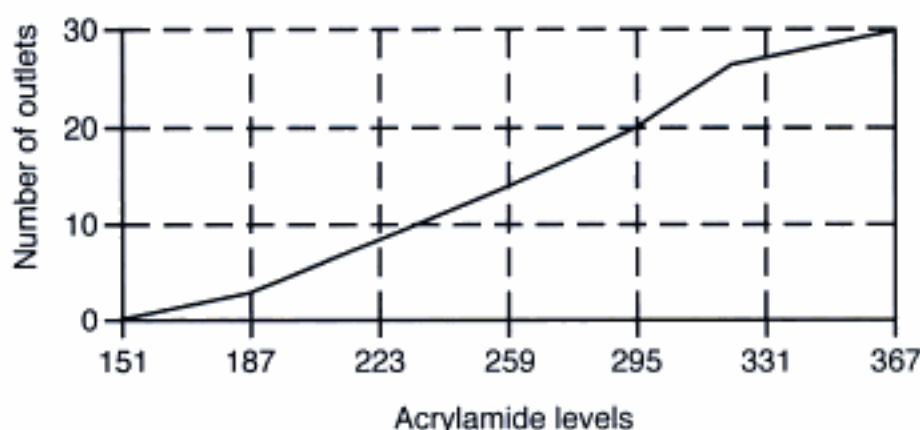
1. The frequency distribution must show class boundaries and cumulative frequencies.
2. The frequency scale on the y -axis must extend to the total of the frequencies.
3. Mark the class boundaries on the x -axis.
4. For each class plot the upper boundary together with the cumulative 'less-than' class frequency.
5. The 'less-than' ogive begins on the horizontal axis ($cum f = 0$) at the lower class boundary of the first class.
6. Draw a smooth curve through the points. The 'less-than' curve slopes upwards and to the right.
7. If the cumulative frequencies are expressed as percentages of the total, a relative ogive can be drawn.

Example 3.14

Below is a frequency table showing the acrylamide levels in the French fries from a sample of Big Mac's outlets. Draw an ogive for this distribution.

Classes	Frequency (f)	x	$cum f$
151 – <187	3	169	3
187 – <223	5	205	8
223 – <259	6	241	14
259 – <295	6	277	20
295 – <331	8	313	28
331 – <367	2	349	30
	30		

Ogive:



Conclusion: Approximately eight of the franchises have acrylamide levels of less than 223. That means that the other 22 outlets have levels above 223.

Activity 3.17

Use your frequency table from activity 3.13 to construct an ogive.

3.4 Using software

There are a number of useful software packages available for data presentation and most of them are simple and easy to use.

Computers can help you develop your ideas about how to organise the information by using a 'try and refine' approach, which would take too long to carry out manually. For example, if you decide to break the information down in a certain way and the results are not what you need, it is a simple matter to create new ways and experiment again.

Computer software can produce accurate and professional graphs and charts from data, but these are only as useful as the data and instructions used to make them.

3.5 Using visual aids

When you are giving a verbal presentation to a group of people, you need to think carefully about the best method of presenting information. The secret of using visual aids successfully is to keep them clear and simple.

Using visual aids in presentations:

- gives more impact to the spoken word
- helps the audience to remember the main points
- gives the audience something to look at other than the presenter.

The visual aids most commonly used to present statistical data are:

- overhead projector (OHP) transparencies
- slides
- flipcharts
- handouts.

TEST YOURSELF 3

1. A questionnaire about how people get news resulted in the following information from a sample of 25. (N=newspaper; T=television; R=radio; M=magazine)

N	T	N	R	N	T	N	R	N
R	T	M	R	M	M	N	M	
M	N	R	T	R	R	T	M	

Summarise the results in a frequency table and construct a pie chart. Interpret your results.

2. The blood of 25 children was tested to determine their blood types. Construct two simple bar charts to display the data using the frequency in one and the %*f* in the other.

Blood type	Frequency	Percentage
A	5	
B	7	
O	9	
AB	4	

- a) According to the data, which blood type is most common?
 - b) According to the data, which blood type is least common?
3. In South Africa, a background check is done on any applicant who applies for a firearm licence. The given table categorises the licence applicants who were denied because of a criminal history for the period 2006 and 2007.

Criminal history	Frequency 2006	Frequency 2007
Domestic violence	64 800	69 100
Felony offence	254 880	215 000
Drug-related offence	30 240	34 029
Other	82 080	73 124

- Construct a simple bar chart to portray the data for each year.
 - Construct a stacked bar chart to portray the data.
 - Construct a multiple bar chart to portray the data.
 - Comment on your results.
4. A recent newspaper article 'The need to be connected' described the results of a survey of 1 000 adults who were asked about how essential various technologies, including personal computers, cell phones and DVD players, influenced their daily lives. The given table summarises the responses:

Response	PC	Cell phone	DVD player
Cannot live without	47%	42%	18%
Would miss, but could do without	27%	26%	35%
Could definitely live without	26%	32%	47%

Construct a comparative bar graph to portray the responses for the different technologies.

5. In the manufacture of printed circuit boards, finished boards are subjected to a final inspection before they are distributed to customers. The type of defect for each board rejected at this final inspection during a randomly selected day is listed together with the frequency of occurrence:

Type of defect	Defects during morning inspection	Defects during afternoon inspection
Etching	5	7
Lamination	10	7
Plating separation	9	8
Poor electrodes coverage	36	40
Low copper plating	122	130

Construct a comparative bar graph to portray the inspections for the different times of the day.

6. Illustrate the following data by means of a multiple and a stacked bar chart.

Johnson and Co. Analysis of costs (R'000)			
Year	Wages	Raw materials	Overheads
1995	15	10	3
1996	14	12	3
1997	10	11	4
1998	11	8	4

7. Use a pie chart to illustrate the following data.

Waxman and Co. Analysis of employment hours lost	
Hours lost through illness	5 100
Hours lost through holidays	2 150
Hours lost through industrial disputes	8 750
Total	16 000

8. A survey of 3 000 adults asked 'How accurate are the weather forecasts in your area?' The responses are summarised in the given table:

Extremely accurate	5%
Very accurate	26%
Sometimes accurate	55%
Not too accurate	9%
Not at all accurate	4%
Not sure	1%

- Construct a pie graph to portray the data.
 - Construct a bar graph to portray the data.
 - Comment on your answers.
9. The volumes of water (in litres) consumed by 24 elephants in one day are listed below:

66 90 68 94 86 96 70 138 90 120 92 102
 82 120 132 82 64 80 88 78 92 66 106 106

Summarise the data set using a stem-and-leaf plot and comment on your results.

10. The number of persons who volunteered to donate blood at a shopping centre was recorded for each of 20 successive Saturdays. The data are shown below:

250 325 333 368 301 386 295 308 320 315
310 332 270 334 356 315 334 370 274 260

Construct a dot plot, stem-and-leaf plot and a frequency distribution for the data.

11. An ecologist wishes to investigate the level of mercury pollution in a stream in the Dullstroom area. He catches 25 trout and measures the concentration of mercury (measured in parts per million) in each fish:

2.2 1.4 1.7 3.4 2.7 2.6 3.0 3.6 3.5 2.6 1.9 3.0 3.8
2.2 2.9 1.8 3.0 3.4 2.8 3.3 3.1 3.2 2.3 2.4 3.7

Construct a dot plot and a stem-and-leaf plot from the data. (Hint: split the stems in two.)

12. The Food and Health Department conducted a study on the calorie content of different types of beer. The calorie content (calories per 100 ml) for 25 different brands of beer is listed below:

43 29 29 31 35 39 42 41 31 27 31 33 32
34 28 22 30 40 33 19 30 32 31 28 23

Construct a stem-and-leaf plot and comment on the calorie content of beer.

Summarise the data in the following data sets (questions 13–28) using a frequency distribution and portray the data using frequency distribution graphs.

13. The following table lists the cholesterol levels from a sample of 65 participants in a research study:

7.8 4.6 6.7 4.6 4.4 6.2 6.0 5.6 5.6 5.4 5.4 4.2 4.5
5.4 4.2 5.2 5.0 5.3 5.3 5.3 5.1 5.1 5.1 5.1 6.6 6.0
5.2 5.0 5.0 4.6 4.8 4.5 4.6 4.0 4.6 3.8 3.3 4.8 4.5
4.8 4.6 4.3 5.0 4.6 4.8 7.1 4.2 4.1 4.0 3.7 4.4 4.0
4.8 4.7 4.2 4.7 4.6 4.3 5.5 4.3 5.0 4.4 4.6 4.4

14. The treatment times (in minutes) for a sample of 50 patients at a health clinic are as follows:

10	12	22	21	20	24	20	35	31	24
24	45	12	26	17	29	19	7	27	29
17	18	16	13	2	16	12	15	22	11
11	15	41	29	16	21	24	14	24	16
8	33	18	21	12	13	15	21	10	33

The health clinic advertises that 90% of all its patients have a treatment time of 40 minutes or less. Do the sample data support this claim? (Hint: Use the cumulative relative ogive to answer the question.)

15. Researchers have investigated lead absorption in children of parents who worked in a factory where lead is used to make batteries. The following shows the levels of lead in the children's blood (in ug/dl of whole blood):

27	25	49	23	31	44	18	37	39	23	13
35	22	14	21	39	17	16	20	15	10	45
24	34	38	48	73	41	35	43	36	34	62

16. The amount of protein (g) for a variety of fast-food sandwiches is reported here:

29	33	23	25	27	40	30	15	35	35	20	18	26
38	27	27	43	57	44	19	35	22	26	22	14	42
35	12	24	24	20	26	12	21	29	34	23	31	15

17. The following are the number of kilometres (in thousands) driven during the year by 110 food inspectors:

40	29	35	33	88	24	38	28	20	21
43	31	18	67	29	76	26	30	23	18
49	44	97	40	48	15	37	43	36	22
55	54	41	34	35	24	38	47	66	34
65	60	32	56	68	38	42	62	55	42
73	31	31	30	36	61	45	52	50	90
30	50	75	20	34	71	51	48	45	84
36	27	52	39	44	51	11	35	41	73
32	65	40	32	81	42	42	53	45	61
10	41	46	84	28	39	47	63	50	52
26	93	36	38	44	58	52	41	55	48

18. A study on the effects of television on the behaviour in adolescents uses, as part of the study, the number of hours per day that the television set is turned on in a household. The following results were obtained in a sample of 30 households:

6.9	4.6	4.3	5.0	6.0	5.3	4.6	3.9	6.0	3.9
6.3	4.2	6.0	5.6	4.2	4.6	6.0	4.3	3.6	6.0
6.0	5.8	3.9	5.7	6.0	3.9	3.7	3.9	3.7	3.9

19. The following data represent tonnes of maize harvested each year for 40 years from Section 20 on an agricultural experiment farm in the Delmas area:

2.71	2.82	1.35	2.20	1.47	2.39	0.59	0.46	1.31	2.50
1.80	0.89	1.64	1.62	1.39	2.19	1.18	1.26	2.04	2.33
1.32	2.60	2.07	0.94	1.42	1.19	2.34	0.77	0.89	1.44
1.62	2.15	0.95	2.02	1.67	1.99	1.48	0.70	0.98	2.00

20. Many people consider the number of calories in an ice-cream bar more important than cost. To investigate the calorie content, a sample of 26 bars gave the following results:

342	310	131	294	209	319	111	353	201	295	182	233	323
234	197	377	439	151	286	147	377	190	182	151	260	301

21. The time, in minutes, for a sample of 70 workers waiting at various points in the production line were as follows:

1	3	7	23	1	2	5	1	0	6
5	0	1	2	4	5	18	0	1	3
0	6	7	1	19	3	5	1	17	3
1	3	8	5	4	14	15	12	0	2
2	5	9	6	11	15	13	17	2	20
1	3	5	16	10	2	5	6	4	4
2	14	5	3	5	6	3	1	11	21

22. From a sample of 36 full-time students, the following information was obtained on the time, in hours, each one spent studying last week:

22	11	33	10	28	7	12	25	32
22	46	21	10	18	17	29	14	2
37	35	3	5	18	4	29	21	20
44	23	31	31	24	13	23	10	36

23. Each of the following figures represents the weight of a package passing through a sorting office. Construct a frequency distribution with cumulative and relative columns:

7.9 7.8 5.0 8.6 8.1 7.9 8.2 8.1 7.3 8.0 8.2 4.9 8.0 7.5 7.4
 8.0 8.0 7.7 7.8 7.5 7.8 5.3 7.9 6.8 7.5 6.9 5.2 8.5 7.9 7.5
 5.2 8.2 4.9 8.7 7.7 7.8 6.0 8.1 8.5 8.0 6.1 7.8 8.1 7.6 7.8
 7.9 7.9 5.3 7.9 8.1 7.6 7.9 8.3 7.4 8.4 7.6 8.0 8.0 8.2 8.2
 6.9 8.1 5.7 7.9 7.7 7.9 6.8 7.8 7.7 7.5 8.1 8.1 8.0 5.1 5.7
 6.0 8.0 5.6 8.2 7.6 7.9 6.2 5.4 5.9 7.8 8.7 6.6 8.1 7.7 6.1
 7.8 7.4 8.1 7.3 7.1

24. The following data were recorded on a study of flexural strength of high-performance concrete obtained by using certain binders and super plasticisers:

9.0 7.9 6.3 7.0 6.5 6.8 7.6 7.0 6.8
 8.1 6.3 7.3 7.2 5.9 8.2 8.7 7.8 9.7
 7.4 7.7 9.7 7.8 7.7 11.6 11.3 11.8 10.7

25. The following observations represent the lifetimes (hours) of a certain type of energy-saver lamp:

612 1 016 1 022 1 003 1 201 883 898 1 029 1 088 1 135
 623 666 744 983 1 029 1 058 1 085 1 122 970 964

26. The following observations were measurements on coating thickness for a sample of low viscosity paint:

0.83 0.88 0.88 1.04 1.09 1.12 1.29
 1.31 1.48 1.49 1.59 1.62 1.65 1.71

27. The following observations are carbon monoxide levels (ppm) in air samples obtained from a certain region in Gauteng:

9.3 10.7 8.5 9.6 12.2 16.6 9.2 10.5
 7.9 13.2 11.0 8.8 13.7 12.1 9.8

28. The age distribution of guards hired by the Jumbo Department within the last year is as follows:

Age	Number
20 - <30	2
31 - <40	13
41 - <50	20
51 - <60	12
61 - <70	3

- Draw the ogives.
 - Based on the graph, the age of about half the guards hired was less than ...
 - Based on the graph, the age of about 30 of the guards hired was more than ...
 - How many of the guards were younger than 35 years old?
 - How many of the guards were older than 50 years old?
-

UNIT 4

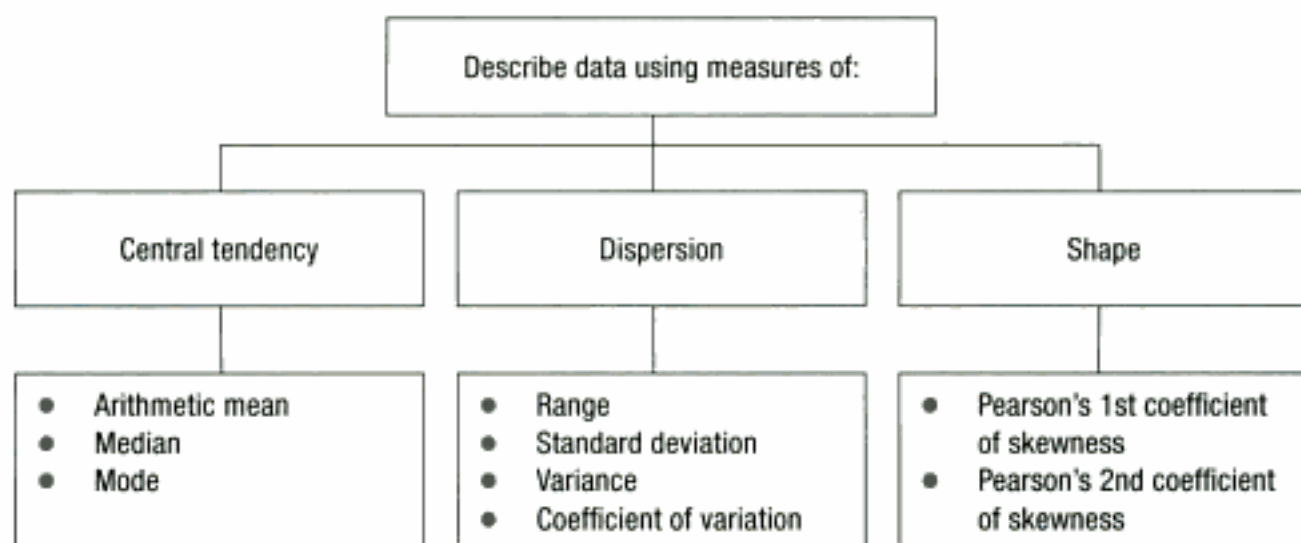
Summarising data using numerical descriptors

In this unit we look at numerical measures that can be used to describe the characteristics of data collected in their raw form (ungrouped data) as well as for data summarised into frequency distributions (grouped data).

After completion of this unit you will be able to:

- compute the mean, median and mode for both grouped and ungrouped data
- describe the characteristics of the mean, median and mode
- compute the range, mean average deviation and standard deviation
- compute and interpret the coefficient of variation and the coefficient of skewness
- locate and plot the mean, median and mode for symmetrical and skewed distributions
- compute measures of relative standing.

Numbers used to describe data sets are called descriptive measures. **Statistics** are summary measures used to describe a sample and populations are described by **parameters**. For the purpose of this text, samples' statistics are calculated and used in later chapters to estimate the population parameters. Data have three major characteristics: location, dispersion and shape.



4.1 Measures of central tendency

Measures of central tendency numerically describe the average or typical value of a dataset. This is a single value that represents the whole data set.

There are many averages, each having its own characteristics. For the same set of data all the averages might have different values. Three commonly used averages are:

1. the arithmetic mean
2. the median
3. the mode.

4.1.1 Arithmetic mean

This is the most commonly used measure of central tendency and is often referred to as the average or the mean. The sample statistic, the mean, is represented by the symbol \bar{x} (x -bar), and the population parameter, the mean, is represented by the Greek letter "mu" (μ).

The arithmetic mean is the sum of all the values in a data set, divided by the number of observations.

The mean of a data set can be seen as the centre of gravity. It is the middle of the actual numerical **values** of all the observations, not necessarily in the middle of the **number** of observations.

Ungrouped data

Ungrouped (or raw) data will usually be presented as a list of numbers in any order or quantity.

$$\bar{x} = \frac{\sum x}{n}$$

Where:

\bar{x} = arithmetic mean

x = each observation value

n = number of observations

Steps

1. Add the values of the individual observations ($\sum x$).
2. Count the number of observations (n).
3. Substitute the totals into the formula for \bar{x} .
4. Divide the sum of the values by the number of observations (n) to obtain \bar{x} .
5. Interpret your answer.

Example 4.1

Calculate the arithmetic mean for the number of cars entering a parking area during a sample of 10-minute intervals.

10 22 31 9 24 27 29 9 23 12

1. Add the numbers:

$$\Sigma x = 10 + 22 + 31 + 9 + 24 + 27 + 29 + 9 + 23 + 12 = 196$$

2. Count the numbers: $n = 10$

3. Use the formula to calculate the mean:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{196}{10} = 19.6 \text{ cars} \approx 20 \text{ cars}$$

We can conclude that on average 20 cars enter the parking area during a 10-minute interval.

Activity 4.1

A city planner working on bikeways needs information about local bicycle commuters. She designs a questionnaire. One of the questions asks how many minutes it takes the rider to pedal from home to his or her destination. A sample of 12 local bicycle commuters yielded the following times:

22 29 27 30 12 22 31 15 26 16 48 23

Determine the mean travelling time.

Grouped data

We cannot calculate exact values of the mean without raw data. If we use grouped data from a frequency distribution, the mean can be approximated using the technique in this section.

The assumption that is necessary is to treat each observation of a class as though it falls on the midpoint (x) of that class. That means that the observations in a particular interval all take the same value.

$$\bar{x} = \frac{\Sigma xf}{n}$$

Where:

x = class midpoint of each class

f = frequency of each class

n = number of observations in the sample (Σf)

Steps

1. Determine the class midpoint (x) of each class.
2. Multiply the midpoint by the frequency to obtain (xf) of each class. Write the products in a column with the heading xf .
3. Sum the xf column to obtain Σxf
4. Sum the frequency column to obtain n : $n = \Sigma f$
5. Substitute the column totals into the formula.
6. Calculate the mean (\bar{x}) for grouped data.
7. Interpret your answer.

Example 4.2

The following frequency table shows the time (in minutes) taken to travel to work for a sample of 25 people from Gauteng. Calculate the mean time to travel to work.

Class boundaries	f	x	xf
15.5 – <21.5	2	18.5	37
21.5 – <27.5	6	24.5	147
27.5 – <33.5	8	30.5	244
33.5 – <39.5	4	36.5	146
39.5 – <45.5	4	42.5	170
45.5 – <51.5	1	48.5	48.5
Total	25	—	792.5

Steps

1. Calculate the midpoints (x) column by adding the lower boundary to the upper boundary of each class and divide the sum by 2.
2. Multiply the midpoint with the frequency of each class to obtain the xf column.
3. Sum the xf column.
4. Sum the f column to obtain n .
5. Substitute the Σxf and n into the formula and calculate the \bar{x} .

$$\begin{aligned}\bar{x} &= \frac{\Sigma xf}{n} \\ &= \frac{792.5}{25} = 31.7 \text{ mm}\end{aligned}$$

The approximate mean time to travel to work for the people of Gauteng is 31.7 minutes.

Activity 4.2

Calculate the mean number of hours of personal computer usage per week for a sample of 16 people and interpret your answer.

Class intervals	<i>f</i>	<i>x</i>	<i>xf</i>
1.95 – <3.95	2		
3.95 – <5.95	5		
5.95 – <7.95	5		
7.95 – <9.95	3		
9.95 – <11.95	1		
Total	16		

Characteristics of the arithmetic mean

1. It is the arithmetic average of all the quantitative measurements in the data set.
2. Every numerical data set has only one mean.
3. It is reliable because it reflects all the values in the data set.
4. It is sensitive to every value in the data set and can be greatly affected by the presence of even a single extreme value (or outlier).

Note: An *outlier* is an unusually large or small observation in comparison with the rest of the values in the data set.

5. It is useful for further inferential statistical procedures.
6. It can be calculated using a pocket calculator with preprogrammed formulae.

4.1.2 Median

The median is the value that occupies the middle position in a data set when arranged in a numerical order. This means that there are an equal number of data values in the ordered distribution that are above it and below it.

Ungrouped data

Steps

1. Arrange the data in a numerical order.
2. Count the number of observations (*n*).
3. Determine the position of the median.

$$\text{median position} = \frac{n + 1}{2}$$

4. Read the value of the median from the number list.
 - If the number of observations is odd, then the median is the value that is exactly in the middle of the data set.

- If the number of observations is even, then the median is the average of the two middle observations in the data set.

Example 4.3

Find the median of each data set.

1. Over a seven-day period, the number of customers (per day) purchasing at Hides Leather Shop was as follows:

4 80 50 10 60 12 5

- Arrange the data in a numerical order:

4 5 10 12 50 60 80

- Determine the position of the median: $\frac{n+1}{2} = \frac{7+1}{2} = \text{value number } 4$
 - Count up to value number 4 on the numerical list: median = 12
 - 50% of the time there were less than 12 customers in the shop and 50% of the time there were more than 12 customers in the shop.
2. A city planner working on bikeways recorded how many minutes it takes the bicycle commuter to pedal from home to his destination. A sample of 12 local bicycle commuters yields the following times:

22 29 27 30 12 22 31 15 26 16 48 23

Determine the median traveling time.

- Numerical order:

12 15 16 22 22 23 26 27 29 30 31 48

- Position of the median: $\frac{n+1}{2} = \frac{12+1}{2} = \text{value number } 6.5$
- Value number 6.5 falls between 23 and 26.
- $\therefore \text{median} = \frac{23+26}{2} = 24.5$
- 50% of the riders took less than 24.5 minutes to destination and 50% took more than 24.5 minutes to travel to destination.

Activity 4.3

1. The following numbers represent the typing speeds of five secretaries in words per minute.

30 90 45 25 55

Determine the median typing speed.

2. How many calories are in a serving of cheese pizza? A variety of pizzas from different outlets were sampled and the calories per serving were determined. The calories were as follows:

332 275 393 347 350 353 357 296 358 322 337 323 333 299

Determine the median calorie content and interpret your answer.

Grouped data

The median for a frequency distribution can be determined either graphically or by calculation.

With grouped data we are unable to determine where the true middle value falls, but we can estimate the median by assuming that the median value will be value number $\frac{n}{2}$ and that the frequencies in the median class are evenly spread. Use the following formula to calculate an estimate for the median.

$$\text{median} = L + \frac{(\frac{n}{2} - F)c}{f_m}$$

Where:

$$n = \Sigma f$$

L = lower boundary of the median class

f_m = frequency of the median class

c = width of interval

F = sum of all the frequencies up to but not including the median class.

Steps

1. Determine the location of the median: $\frac{n}{2}$
2. Construct the cumulative frequency column ($\text{cum} < f$).
3. Compare the position of the median with the $\text{cum} < f$ column to determine which one of the intervals contains the median. The median class is the interval where the cumulative frequency is equal to or exceeds $\frac{n}{2}$ for the first time.
4. Estimate the value of the median using the formula for grouped data.

Example 4.4

Calculate the median time (in minutes) taken to travel to work for a sample of 25 people in Gauteng.

Class boundaries	f	$\text{cum} < f$
15.5 – <21.5	2	2
21.5 – <27.5	6	8
27.5 – <33.5	8	16
33.5 – <39.5	4	20
39.5 – <45.5	4	24
45.5 – <51.5	1	25
Total	25	

1. Calculate the cumulative frequency column ($\text{cum} < f$).
2. Determine the position of the median: $\frac{25}{2} =$ value number 12.5

- Compare 12.5 with the values in the $cum < f$ column. You will see that up to the end of the 2nd class we have 8 values. Therefore value number 12.5 must fall in class number 3 which contains values number 9 to 16.
- The median will fall somewhere between 27.5 and 33.5.
- Substitute the required values from that class into the formula and calculate the median.

The median time to travel to work for the 25 people in the sample is:

$$\begin{aligned} \text{median} &= L + \frac{(\frac{n}{2} - F)c}{f_m} \\ &= 27.5 + \frac{(12.5 - 8)6}{8} \\ &= 30.88 \end{aligned}$$

This means that half the sampled people traveled less than 30.88 minutes to work and the other half more than 30.88 minutes.

Use the ogive to determine the median

You can determine the value of the median graphically by making use of the ogive.

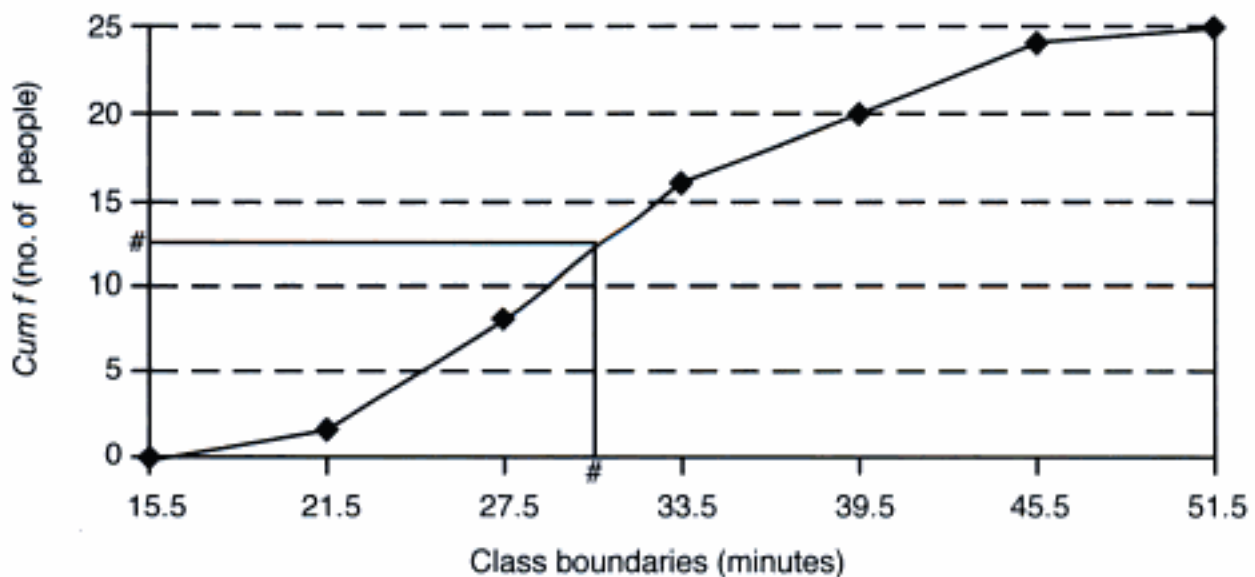
Steps

- Draw the cumulative 'less-than' ogive.
- Find the median position ($\frac{n}{2}$) on the vertical axis.
- Draw a straight horizontal line up to the ogive. Drop a straight line down to the x -axis.
- The corresponding value on the horizontal x -axis is the median value.

Example 4.5

Use the table from example 4.4 and determine the median value graphically.

- Find the position on the vertical axis: $\frac{n}{2} = \frac{25}{2} = 12.5$



2. Draw a straight horizontal line from the position on the y-axis up to the ogive. Drop a straight line down to the x-axis.
3. Read the median value from the x-axis: median = ± 31 minutes.

Activity 4.4

Calculate the median number of hours of personal computer usage per week for a sample of 16 people by making use of a formula as well as a graph.

Class intervals	<i>f</i>	
1.95 – <3.95	2	
3.95 – <5.95	5	
5.95 – <7.95	5	
7.95 – <9.95	3	
9.95 – <11.95	1	
Total	16	

Characteristics of the median

- It is the central value: 50% of the measurements lie above it and 50% lie below it.
- There is only one median for a data set.
- Extreme measurements do not affect the median as strongly as they do the mean because the median is dependent on the middle position of the sample population which excludes outliers at the beginning or end of the ordered data set.
- It is a better measure of central tendency to use than the mean when the data are very skewed.
- Its computation does not involve every measurement in the data set.

4.1.3 Mode

The mode of a data set is the value that occurs most frequently. It can be a good measure to represent a typical value such as the most popular shirt size.

Ungrouped data

Tally the number of observations that occur for each data value. If there is no value that occurs more often than the others, then there is no mode. (**Note:** this is not the same as a mode of zero.) A set of data can have no mode, one mode or more than one mode (bi-modal or multi-modal).

Example 4.6

1. The commission earnings of five colleagues for the previous month was as follows:

R5 000 R5 200 R5 200 R5 700 R8 600

Modal commission was R5 200 because more of the colleagues earn R5 200 than any other income.

2. The lengths of stay (in days) for a sample of nine patients in Ward A are:

17 19 19 4 19 26 4 21 4

The modal lengths of stay are 19 and four days: more of the patients stay either 4 days or 19 days than any other number of days.

3. The hourly income rates of five workers are:

R4 R9 R7 R16 R10

There is no mode: none of the workers earn the same income rate.

Activity 4.5

A telephone company conducted a study on the length of long-distance calls. A sample of 10 calls gave the following lengths in minutes:

1.4 15.5 2.1 8 15.5 1.4 17.7 7.2 9.1 15.5

Determine the modal length and comment on your answer.

Grouped data

An estimate of the mode can be approximated either graphically or by making use of a formula. Grouped data do not show a single most frequently occurring value but assume that the mode will occur in the interval with the highest frequency.

$$\text{mode} = L + \left(\frac{\Delta_1}{\Delta_2 + \Delta_3} \right) c$$

Where:

L = lower boundary of modal class

c = width of interval

Δ_1 = frequency (f) of modal class minus f of previous class

Δ_2 = frequency of modal class minus f of following class

Steps

1. Select the class containing the highest frequency as the modal class.
2. Determine the Δ_1 value by subtracting the frequency of the class preceding the modal class from the frequency of the modal class.
3. Determine Δ_2 by subtracting the frequency of the class following the modal class from the frequency of the modal class.

- Use the formula to estimate the modal value.
- Interpret your answer.

Example 4.7

The modal time (in minutes) taken to travel to work for a sample of 25 people is:

Class boundaries	<i>f</i>
15.5 – <21.5	2
21.5 – <27.5	6
27.5 – <33.5	8
33.5 – <39.5	4
39.5 – <45.5	4
45.5 – <51.5	1
Total	25

- Choose the class with the highest frequency – that is class number 3.
- Mode = $L + \left(\frac{\Delta_1}{\Delta_2 + \Delta_3} \right) c = 27.5 + \left(\frac{2}{2 + 4} \right) 6 = 29.5$ min
- More of the people take 29.5 minutes to travel to work than any other time.

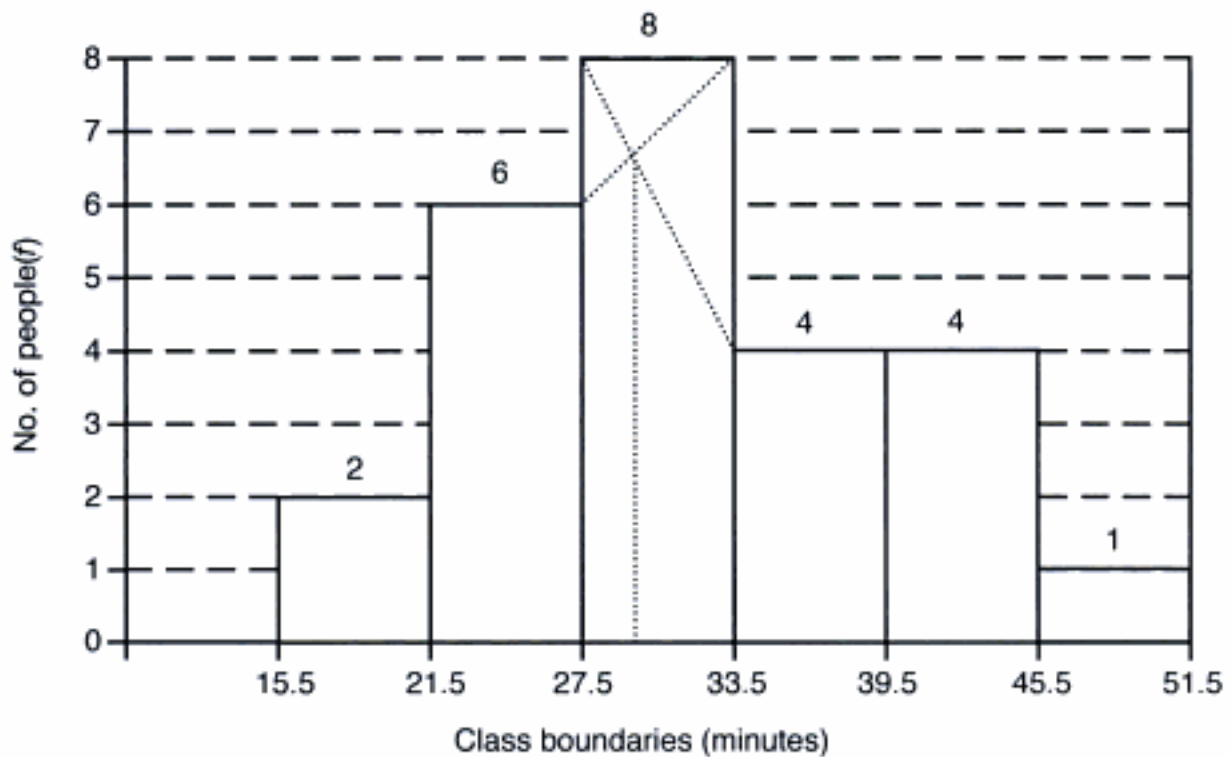
Use the histogram to approximate the mode

Steps

- Draw the histogram of the frequency distribution.
- Identify the longest bar on the histogram as the modal bar.
- Draw a line from the top right corner of the modal bar up to right corner of the bar to its immediate left.
- Draw a second line from the top left corner of the modal bar up to the top left corner of the bar to its immediate right.
- Draw a straight line parallel to the *y*-axis through the intersection point of the previous two lines down to the *x*-axis.
- The value on the *x*-axis approximates the modal value.

Example 4.8

This histogram is constructed from the frequency distribution in example 4.7 and is used to determine the modal time travelled to work graphically.



The mode is ± 29 min.

Activity 4.6

Calculate the modal number of hours of personal computer usage per week for a sample of 16 people using a formula and read the modal number of hours from an appropriate graph. Interpret your result in the context of the data.

Class intervals	f
1.95 - <3.95	2
3.95 - <5.95	5
5.95 - <7.95	5
7.95 - <9.95	3
9.95 - <11.95	1
Total	16

Characteristics of the mode

- It is the most frequent or probable measure in the data set.
- It is not based on all the data in the sample.
- It is not affected by extreme values.
- It can be used for quantitative and qualitative data.
- Sometimes the mode does not exist or it is possible that there is more than one mode. For these reasons, many people consider the mode as an unreliable measure of central tendency.

4.1.4 Choose between the mean, median or mode

The whole point of using an average is that it should convey an impression of a distribution in a single value. It is important to use the right type of average. All the averages are different measures with different uses. The factors that play a role in choosing the right average are the following:

1. Is the nature of the data numerical or non-numerical?
 - The **mode**, which is the value that occurs most often, is the only measure of central tendency useful for qualitative (non-numerical) data that you cannot rank in any way. You can also use the mode for all other qualitative or quantitative (numerical) data sets.
 - If you can rank qualitative data sets, you can use the **median**. The median is also valid for all quantitative data sets.
 - The **arithmetic mean** involves arithmetic and is appropriate only for quantitative data sets.
2. What does each average tells us?

Depending on the situation and the problem under investigation, one measure may be superior to another, and in some other cases, you can use all three in conjunction.

 - The **mode** identifies the most common or 'typical' value, or the value that occurs more often than the others do. It may be a good choice if one value occurs *much* more often than others do. At the same time, the mode conveys the least amount of information about the data set as a whole. In some samples, the mode may be in the middle of the distribution but in others, it may be a value at one end of the distribution. It is also possible to have more than one mode that will eliminate the mode as an option. Outliers do not influence the mode at all and stays at the peak of the distribution.
 - The **median** indicates the centre of the distribution. There are the same number of observations that lie above and below the median, regardless of how far above or how far below. This means that it is unlikely that outliers at either end of the distribution will affect the median very much.
 - The **mean** is the most frequently used average because it includes all the values in the data set. This feature makes it the most sensitive to extreme values.
3. What is the shape of the distribution?
 - In a symmetrical distribution, the mean, median and mode will be the same or very close together. Whichever one you choose will give you the same answer.

- If there are extreme values present on one side of the data set, the distribution is skewed. If the mean is very different from the median, the median will be a better option to use.
- Skewness will be discussed later in the unit.

Example 4.9

1. The test marks of five students are as follows:

55 59 66 66 94

The arithmetic mean of the marks is 68. This means that the sum of all the marks evenly divided by all the learners will give you 68. The median value is 66, which means that half of the learners scored less than 66 and the other half scored more than 66. The mode is 66, which means that more learners obtained 66 than any other mark.

If the values are arranged in a numerical order and you slot the arithmetic mean value in position, you will see that there are 4 values smaller than the mean and only one bigger than the mean. This means that the value on the right is an outlier which pulled the mean to the right, causing the distribution to be positively skewed. For this reason the median or the mode will be a better measure to choose.

2. A student obtained the following 4 marks in mathematics:

88 75 95 100

The arithmetic mean of 89.5 would most probably be the best average to use since it takes into account all the test marks of one student and therefore indicates overall performance.

3. In calculating the average house prices in a particular suburb in Gauteng, you will most probably make use of the median. This is because the relatively few homes with extremely high or low prices do not affect the median strongly. The median provides a better indication of the average house price.

Activity 4.7

The National Housing Department conducted a survey to estimate the average number of livable square meters for low cost housing. The reported mean was 24.5 square meters and the median 22.2 square meters. Which measure of central tendency is more appropriate? Explain your answer.

4.2 Measures of dispersion

An average summarises a set of data in just one number. Two sets of data can have the same mean value and yet be very different if one is more spread out than the other. To describe this difference quantitatively, we use a measure of dispersion. This is a descriptive measure that indicates the amount of variation in a data set. Some commonly used measures of dispersion are the range, mean absolute deviation, standard deviation and variance.

4.2.1 The range

The range is the difference between the largest and smallest values in a data set. It measures the distance across the entire set of data but its usefulness as a measure of dispersion is limited. The range tells the difference between the largest and smallest values in the distribution, it does not tell us how much other values vary from one another or from the mean.

$$\text{range} = \text{maximum value} - \text{minimum value}$$

For grouped data the range is the difference between the upper boundary of the last interval and the lower boundary of the first interval.

Example 4.10

A bakery regularly orders punnets of blueberries for its famous blueberry cheese cake. The average weight of the punnets is supposed to be 600 g. The baker uses one punnet of blueberries in each cake. It is important that the punnets are of consistent weight so that the cake turns out right. Random samples of punnets from two suppliers were weighed. The weights in grams of the punnets were:

Supplier 1: 480 600 600 600 760

Supplier 2: 480 540 570 760 760

Calculate the range of punnet weights for each supplier and comment on your results:

Supplier 1: $\text{range} = 760 - 480 = 280 \text{ gr.}$

Supplier 2: $\text{range} = 760 - 480 = 280 \text{ gr.}$

The ranges are the same, but it is obvious that the variations within the samples are different. So the range will not solve the bakery's problem if they want to choose the supplier that will provide punnets with consistent weights.

4.2.2 Mean absolute deviation (MAD)

This is a better measure of dispersion than the range because it takes every observation into account and measures variability around the average; it measures how much the data differs from the mean.

Note: The deviation of a measurement in a data set is the difference between the entry and the mean of the data set.

Some of the measurements are smaller than the mean, which will result in a negative deviation and others are larger than the mean, which will result in a positive deviation.

To prevent negative deviations from the mean cancelling positive deviations, the algebraic signs of the deviations are ignored and the absolute differences are averaged.

Ungrouped data

Steps

1. Calculate the arithmetic mean (\bar{x}) of the distribution
2. Determine the difference between each value and \bar{x} without regard to the algebraic sign: $|x - \bar{x}|$. The two straight lines indicate that you are using the absolute value.
3. Add the absolute values of the deviations: $\sum|x - \bar{x}|$
4. Divide the sum by the number of values (n).

$$\text{MAD} = \frac{\sum|x - \bar{x}|}{n}$$

Example 4.11

Calculate the mean absolute deviation for the number of cars entering a parking area during a sample of 10-minute intervals.

x	$ x - \bar{x} $
10	9.6
22	2.4
31	11.4
9	10.6
24	4.4
27	7.4
29	9.4
9	10.6
23	3.4
12	7.6
$\Sigma x = 196$	76.8

1. $\bar{x} = \frac{196}{10} = 19.6$
2. $\sum|x - \bar{x}| = 76.8$
3. $n = 10$

4. $MAD = \frac{\sum|x - \bar{x}|}{n} = \frac{76.8}{10} = 7.68$ cars
5. The typical deviation from the mean is 7.68 cars. The smaller the answer, the less variation we have in the distribution.

Activity 4.8

In a study on bikeways the number of minutes it takes a sample of 12 local bicycle commuters to pedal from home to their destination is recorded.

22 29 27 30 12 22 31 15 26 16 48 23

Determine the mean absolute deviation traveling time for the riders and interpret the result in the context of the data.

MAD for grouped data

Steps

1. Calculate the arithmetic mean (\bar{x}) of the distribution.
2. Determine the deviation of each midpoint (x) from \bar{x} without regard to the algebraic sign: $|x - \bar{x}|$
3. Multiply the absolute deviation in each class by the frequency of that class. $|x - \bar{x}|f$
4. Add the absolute values of the deviations: $\sum|x - \bar{x}|f$
5. Divide the sum by the number of values ($n = \sum f$)

$$MAD = \frac{\sum|x - \bar{x}|f}{n}$$

Example 4.12

The following frequency table shows the time (in minutes) taken to travel to work for a sample of 25 people from Gauteng. Calculate the mean absolute deviation time to travel to work.

Class boundaries	f	x	$ x - \bar{x} f$
15.5 – <21.5	2	18.5	26.40
21.5 – <27.5	6	24.5	43.20
27.5 – <33.5	8	30.5	9.60
33.5 – <39.5	4	36.5	19.20
39.5 – <45.5	4	42.5	43.20
45.5 – <51.5	1	48.5	16.80
Total	25		158.40

$$1. \quad \bar{x} = \frac{\Sigma xf}{n} = \frac{792.5}{25} = 31.7 \text{ min}$$

2. Construct the $\Sigma|x - \bar{x}|f$ column and sum the results.

3. Substitute the answers into the formula to determine the MAD.

$$\text{MAD} = \frac{\Sigma|x - \bar{x}|f}{n} = \frac{158.4}{25} = 6.34 \text{ minutes}$$

The average absolute difference between each observation of the time taken to travel to work and the mean is 6.34 minutes.

Activity 4.9

The following frequency distribution summarises the number of hours of personal computer usage per week for a sample of 16 people. Calculate the mean absolute deviation for the sample time.

Class intervals	<i>f</i>
1.95 - <3.95	2
3.95 - <5.95	5
5.95 - <7.95	5
7.95 - <9.95	3
9.95 - <11.95	1
Total	16

4.2.3 Standard deviation (*s*)

The standard deviation is the most widely used measure of dispersion and measures on the average, how far each data value is from the mean. To prevent negative deviations from the mean cancelling positive deviations, the differences are squared.

1. It uses all the entries in the data set and is therefore sensitive to outliers.
2. The larger the standard deviation the larger the variation in the data. A standard deviation of zero means there is no variation.
3. It is useful for further inferential statistical procedures because most statistical theories are based on distributions described by their mean and standard deviation.
4. The measuring unit is expressed in the original units of measurements. (rands, minutes, metres, etc.).
5. It can be calculated using a pocket calculator with preprogrammed formulae.

Ungrouped data

Steps

1. Compute the arithmetic mean (\bar{x}).
2. Find the deviation of each observation by subtracting \bar{x} from each data value: $(x - \bar{x})$
3. Square each difference: $(x - \bar{x})^2$
4. Sum the squared differences: $\Sigma(x - \bar{x})^2$
5. Calculate the average by dividing the sum by $(n - 1)$.

Note: Division by $(n - 1)$, known as degrees of freedom, is to correct the bias in estimating the population standard deviation using the sample standard deviation.

6. The standard deviation is the square root of this total.

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

7. A large amount of variability in the sample is indicated by a relatively large value of the standard deviation, whereas a standard deviation close to zero indicates a small amount of variability. The standard deviation can be interpreted as a 'typical' deviation from the mean. If two samples are compared, we can say that the sample with the smaller standard deviation has less variability than the one with the higher standard deviation.

Example 4.13

Calculate the standard deviation for the number of cars entering a parking area during a sample of 10-minute intervals.

x	$(x - \bar{x})$	$(x - \bar{x})^2$
10	-9.6	92.16
22	2.4	5.76
31	11.4	129.96
9	-10.6	112.36
24	4.4	19.36
27	7.4	54.76
29	9.4	88.36
9	-10.6	112.36
23	3.4	11.56
12	-7.6	57.76
$\Sigma x = 196$	0.0	684.4

- $\bar{x} = \frac{196}{10} = 19.6$.
- Construct a $(x - \bar{x})$ column: $\Sigma(x - \bar{x}) = 0$.
- Square each deviation in the $(x - \bar{x})$ column to obtain $(x - \bar{x})^2$.
- Sum the squared differences column to obtain: $\Sigma(x - \bar{x})^2 = 684.4$
- Divide the sum by $(n - 1)$. There are 10 observations in the sample, therefore:
 $10 - 1 = 9$
- Substitute the calculated sums into the formula to determine the standard deviation.
- $s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}} = \sqrt{\frac{684.4}{10 - 1}} = 8.72$
- A typical deviation for the number of cars is 8.72.

Activity 4.10

In a study on bikeways the number of minutes it takes a sample of 12 local bicycle commuters to pedal from home to their destination is recorded.

22 29 27 30 12 22 31 15 26 16 48 23

Determine the standard deviation of the traveling time for the riders and interpret the result in the context of the data.

Grouped data

To estimate the standard deviation from data grouped into a frequency distribution, we assume that each class is represented by its midpoint (x).

Steps

- You need a frequency table with the following columns: classes, frequencies and midpoints.
- Compute the arithmetic mean (\bar{x})
- Subtract the mean from each class midpoint and square the difference: $(x - \bar{x})^2$
- Multiply the squared difference by the frequency within each class: $(x - \bar{x})^2 f$
- Sum the result to obtain the total squared deviation from the mean: $\Sigma(x - \bar{x})^2 f$
- Calculate the average of this total by dividing by $(n - 1)$
- The standard deviation is the square root of this total.

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2 f}{n - 1}}$$

Example 4.14

The following frequency table shows the time (in minutes) taken to travel to work for a sample of 25 people from Gauteng. Calculate the standard deviation of the time to travel to work.

Class boundaries	f	x	$(x - \bar{x})^2 f$
15.5 - <21.5	2	18.5	348.48
21.5 - <27.5	6	24.5	311.04
27.5 - <33.5	8	30.5	11.52
33.5 - <39.5	4	36.5	92.16
39.5 - <45.5	4	42.5	466.56
45.5 - <51.5	1	48.5	282.24
Total	25	—	1512

- $\bar{x} = \frac{\Sigma xf}{n} = \frac{792.5}{25} = 31.7$ min (from example 4.2).
- Subtract 31.7 from each x -value and square the difference.
- Multiply each squared difference by the frequency of that difference and record the answers in the $(x - \bar{x})^2 f$ column. The first value in this column is calculated as $(18.5 - 31.7)^2 \times 2$.
- Sum the results.
- Substitute the total into the formula to determine the standard deviation.

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2 f}{n - 1}} = \sqrt{\frac{1512}{25 - 1}} = 7.94 \text{ minutes}$$

The typical standard deviation between each observation of traveling time and the mean is 7.94 minutes.

Activity 4.11

The following frequency distribution summarises the number of hours of personal computer usage per week for a sample of 16 people. Calculate the standard deviation for the sample times.

Class intervals	<i>f</i>
1.95 - <3.95	2
3.95 - <5.95	5
5.95 - <7.95	5
7.95 - <9.95	3
9.95 - <11.95	1
Total	16

4.2.4 Variance (s^2)

The variance is the standard deviation squared (s^2). Although this is a very popular measure in describing data, the main drawback is that the unit of measure is also squared. Statistics measured in squared units are problematic to interpret. If the standard deviation is equal to 5.25 hours, the variance will be 27.56 hours squared.

4.2.5 Coefficient of variation (CV)

The coefficient of variation is a relative measure of dispersion, which is the ratio, expressed as a percentage, of the standard deviation to the mean. This is sometimes used as measure of risk.

$$CV = \frac{s}{\bar{x}} \times 100$$

This is a unit-free number because the standard deviation and mean are measured using the same units. The higher the result the more variability there is in a set of data.

All the measures of dispersion described so far have dealt with a single set of data. In practice, it is often important to compare two or more sets of data with different means, sample sizes or measurement units.

Example 4.15

A manufacturing company produces a product in two sizes: a 1 000 ml bottle and a 500ml bottle. Because of mechanical variability in the filling machinery, there is a standard deviation of 50 ml and 40 ml respectively. The machine with the lowest CV will be the more consistent.

$$CV(1\,000\text{ ml}) = \frac{50}{1\,000} \times 100 = 5\%$$

$$CV(500\text{ ml}) = \frac{40}{500} \times 100 = 8\%$$

For the 1 000 ml bottle, the CV of the filling process is 5% of the filling mean. For the 500 ml bottle, the CV is 8% of the filling mean.

Although the machine filling the smaller bottle has a lower standard deviation, the CVs indicate that it is the machine filling the larger bottle which is relatively more consistent.

Activity 4.12

Two growers of grapefruit have obtained the following statistics regarding the mass of their current crops.

Grower A: $\bar{x} = 300$ g with $s = 20$ g

Grower B: $\bar{x} = 280$ g with $s = 40$ g

Which grower's grapefruit are more uniform in mass?

4.3 Measures of shape

Measures of shape are tools that can be used to describe the shape of a distribution. Two measures of shape are:

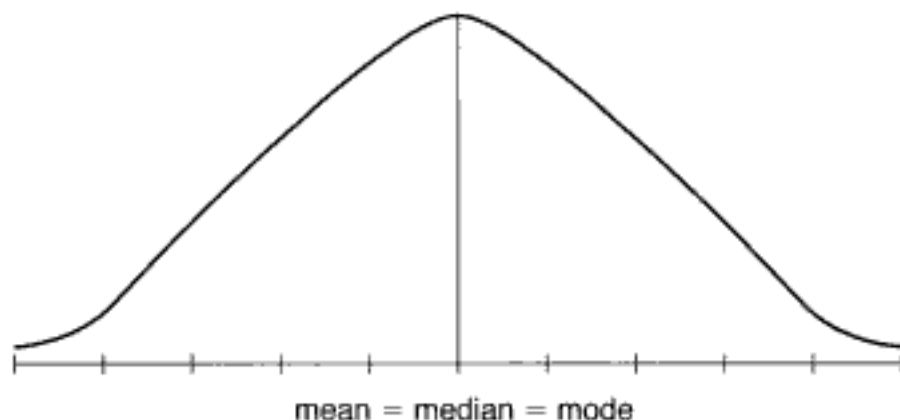
1. Skewness, which measures its symmetry or lack of symmetry.
2. Kurtosis, which measures its peakedness.

4.3.1 Skewness (SK)

You can describe the shape of a distribution by its symmetry or lack thereof (skewness), which relates to the shape of the histogram, polygon, stem-and-leaf or dot plot that you can draw from the data. The shape influences the locations of the mean, median and mode in the data set, for example whether the mean is larger or smaller than the median.

In *symmetrical* or *normal distributions*, the left half is a mirror image of the right. When a symmetrical distribution has a single mode, the mode will be in the center of the distribution. Furthermore, the mean and the median will be equal to the mode. There are no extreme values on the one side to pull the mean away from the bulk of the data. The skewness coefficient will take a zero value.

To portray the shape of a distribution, you can make use of the histogram or a smooth polygon.



*image
not
available*

Pearson's second coefficient of skewness

This coefficient compares the mean and median in context of the magnitude of the standard deviation.

$$SK_2 = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

Simply by knowing the value of the skewness coefficient, we can infer the general shape of the distribution without resorting to a diagram:

- If $SK_2 = 0$, the distribution is symmetrical with the mean = median = mode.
- If $SK_2 > 0$ but less than 3, the distribution is positively skewed with the mean $>$ median $>$ mode.
- If $SK_2 < 0$ but greater than -3 the distribution is negatively skewed with the mean $<$ median $<$ mode

Example 4.16

If the time taken to complete a particular complex task resulted in a mean of $\bar{x} = 34.34$ minutes, a median = 35.27 minutes and a $s = 6.86$, calculate SK.

$$SK_2 = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}} = \frac{3(34.34 - 35.27)}{6.86} = -0.14$$

The distribution is negatively skewed but close to normal.

Activity 4.13

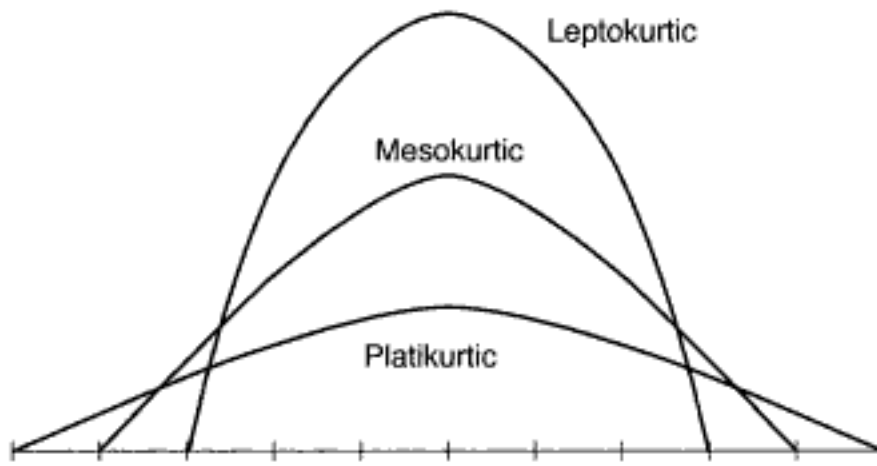
In a sample showing the sodium contents (in milligrams per kilogram) of chocolate pudding made from instant mix, the mean $\bar{x} = 2\,965.2$, the median = 2 946 and the standard deviation = 543.52

Calculate the coefficient of skewness for this distribution and interpret your answer. Illustrate your answer with a rough sketch.

4.3.2 Measures of kurtosis

Kurtosis describes the amount of peakedness of a distribution. The flatter the curve, the greater the spread of the data. This means that the standard deviation is larger relative to the mean. Although a formula exists to measure kurtosis, it is easier to determine the extent of the kurtosis by observing the frequency curve or polygon.

- A distribution that is high and thin is referred to as leptokurtic.
- A distribution that is flat and spread out is referred to as platykurtic.
- A distribution that is more normal in shape, that is either very peaked nor flat, is referred to as mesokurtic.



4.4 Interpreting centre and variability

1. Dispersion is the amount of spread or scatter that occurs in a data set. It can be interpreted as the size of a 'typical' deviation from the mean. If the values in the data set are clustered tightly about their mean, the standard deviation is small, but if the values are widely dispersed about their mean, the standard deviation is large.
2. In comparing two data sets with the same unit of measure, the one with the larger standard deviation has the greater amount of variability and the one with the smaller standard deviation is more consistent, with less variability among the numbers in the data set.
3. If you have a single data set, the mean can be combined with the standard deviation to obtain information about how values in a data set are distributed along a number line. To do this, we describe how far away a particular observation is from the mean in terms of the standard deviation. For example, we might say that an observation is two standard deviations above the mean or one standard deviation below the mean.

Consider a data set with a mean of 100 and a standard deviation of 15.

- The mean minus one standard deviation = $100 - 15 = 85$. This means that '85' is one standard deviation below the mean.
- $100 + 15 = 115$. This means that '115' is one standard deviation above the mean.
- All observations that fall between '85' and '115' are within one standard deviation from the mean.
- Two standard deviations = $2 \times 15 = 30$.
 $100 - 30 = 70$ and $100 + 30 = 130$. All observations that fall between '70' and '130' are within two standard deviations from the mean.
- $100 + 3(15) = 145$. Observations above 145 exceed the mean by more than three standard deviations.

4. The following two rules can be applied, depending on the shape of the distribution.
- If the distribution is symmetrical, you can make a statement about the proportion of data values that fall into various intervals, using the **empirical rule**.
 - A more general interpretation of the standard deviation is derived from **Chebysheff's theorem**, which applies to distributions of all shapes.

Empirical rule:

- Approximately 68% of all observations fall within one standard deviation from the mean.
- Approximately 95% of all observations fall within two standard deviations from the mean.
- Approximately 99.7% of all observations fall within three standard deviations from the mean.

Chebysheff's theorem:

The proportion of observations in any sample that lies within k standard deviations of the mean is at least $1 - \frac{1}{k^2}$ for $k > 1$

With $k = 2$, Chebesheff's theorem states that at least $1 - \frac{1}{2^2} = \frac{3}{4}$ or 75% of all observations will fall within two standard deviations of the mean. That will be the values between $\bar{x} - 2k$ and $\bar{x} + 2k$

With $k = 3$, Chebesheff's theorem states that at least $1 - \frac{1}{3^2} = \frac{8}{9}$ or 89% of all observations will fall within three standard deviations of the mean. That will be the values between $\bar{x} - 3k$ and $\bar{x} + 3k$

Example 4.17

A psychologist randomly selected 10 TV cartoons and counted the number of incidents of verbal and physical violence in each. The counts were as follows:

13 26 16 21 15 31 15 30 14 11

1. $\bar{x} = 19.2$ hours
2. $s = 7.33$ hours
3. Range: 31 - 11

4. The stem-and-leaf plot shows a positive skewed distribution. Pearson's second coefficient of skewness can also be used to determine the shape.

Stem	Leaf
1	134556
2	16
3	01

5. This distribution is positively skewed therefore you can use Chebesheff's theorem to come to the following conclusions:
- At least 75% of the data points (x) will fall within two standard deviations from the mean: $19.2 + 2(7.33) = 33.86$
 $19.2 - 2(7.33) = 4.54$
That will be the values between 4.54 and 33.86.
 - Within three standard deviations from the mean you will find at least 89% of all the data values falling between -2.79 and 41.19 .
 - As you can see from the range, all the values in the data set fall within two standard deviations from the mean.

Example 4.18

The stem-and-leaf display below displays the IQ scores of a sample of 112 children.

Stem: tens
Leaf: ones

Stem	Leaf
6	1
7	25679
8	0000124555668
9	0000112333446666778889
10	0001122222333566677778899999
11	00001122333344444477778899999
12	01111123445669
13	006
14	26
15	2

The summary statistics for this distribution are: $\bar{x} = 104.5$ and $s = 16.3$

This distribution can be reasonably well described by a symmetrical or normal shape; therefore the empirical rule can be applied.

- Within one standard deviation of the mean:
 $[104.5 - 1(16.3)]$ and $[104.5 + 1(16.3)]$
 $= 88.2$ and 120.8
 Approximately 68% of the IQ scores are between 88.2 and 120.8.
- Within two standard deviations of the mean:
 $[104.5 - 2(16.3)]$ and $[104.5 + 2(16.3)]$
 $= 71.9$ and 137.1
 Approximately 95% of the IQ scores are between 71.9 and 137.1.
- Within three standard deviations of the mean:
 $[104.5 - 3(16.3)]$ and $[104.5 + 3(16.3)]$
 $= 55.6$ and 153.4
 Approximately 99.7% of the IQ scores are between 55.6 and 153.4.

Activity 4.14

The following frequency table shows the time (in minutes) taken to travel to work for a sample of 25 people from Gauteng. Interpret the centre and variability for the sample.

Class boundaries	<i>f</i>
15.5 – <21.5	2
21.5 – <27.5	6
27.5 – <33.5	8
33.5 – <39.5	4
39.5 – <45.5	4
45.5 – <51.5	1
Total	25

Use the histogram or Pearson's second coefficient of skewness to determine if the distribution is close to normal or skewed. From previous examples we know that:

$$\bar{x} = 31.7 \text{ min}$$

$$s = 7.94 \text{ min}$$

$$\text{median} = 30.88 \text{ min}$$

Activity 4.15

Interpret the centre and variability of the number of cars entering a parking area during a sample of ten-minute intervals. From previous examples we know that:

$$\bar{x} = 19.6$$

$$s = 8.72$$

$$\text{median} = 22.5$$

10 22 31 9 24 27 29 9 23 12

4.5 Measures of relative standing

These measures are used to determine the position of an observation in a set of data in relation to the other values in the set. The most familiar measures are quartiles and percentiles, also known as fractiles.

If you want to know what data value in a sample has a certain percentage of the sample data above or below it, you calculate the **percentile**. Percentiles divide the data into 100 equal parts and each percentile (P_j) is a value such that $j\%$ of the observations are smaller. (j can take on a value between 1% and 100%.)

Certain percentiles are used frequently. These are the 25th percentile and the 75th percentile, also known as the first and third quartiles. **Quartiles** divide the data into four equal parts. The first quartile (Q_1) is a value such that 25% of the observations are smaller and the third quartile (Q_3) is a value such that 75% of the values are smaller.

The median, which is a measure of central tendency, is also a measure of relative standing. As you have learned previously, the median divides the data into two equal parts, the bottom 50% and the top 50%. The median is the middle quartile, Q_2 or P_{50} . The exact location must therefore be determined before the value can be calculated.

Ungrouped data

Steps

1. Arrange the numbers in a numerical order.
2. Change quartiles to percentiles ($Q_1 = P_{25}$ and $Q_3 = P_{75}$)
3. Determine the position of the percentile or quartile you want to obtain. (Where in the numerical list will you locate this value?) Use the following formula to determine the position of the quartile or percentile:

$$\text{position } P_j = \frac{jn}{100}$$

With ' j ' the j th percentile and ' n ' the number of observations.

4. If the position results in a fraction, choose the next larger integer. If the position results in an integer, add 0.5.
5. Read the value from the numerical list.

Example 4.19

A psychologist randomly selected 10 TV cartoon shows and counted the number of incidents of verbal and physical violence in each. The counts were as follows:

26 13 16 21 15 31 15 30 14 11

Determine Q_3 , Q_1 , P_{80} and P_{20} .

1. Numerical order:

11 13 14 15 15 16 21 26 30 31

2. Determine the position of each value and read the value from the array.

$$Q_1 \text{ position} = P_{25} = \frac{25(10)}{100} = 2.5 \text{ round to position 3.}$$

$\therefore Q_1$ value = 14. This means that 25% of the TV cartoons have less than 14 incidences of verbal and physical violence per cartoon and the other 75% have more than 14 incidences per cartoon.

$$Q_3 \text{ position} = P_{75} = \frac{75(10)}{100} = 7.5 \text{ round to position 8.}$$

$\therefore Q_3$ value = 26. This means that 75% of the cartoons have less than 26 incidences of verbal and physical violence per cartoon and 25% have more than 26 incidences.

$$P_{80} \text{ position} = \frac{80(10)}{100} = 8 \text{ round to position 8.5.}$$

$$\therefore P_{80} \text{ value} = \frac{26 + 30}{2} = 28$$

This means that 80% of the cartoons have less than 28 incidences and 20% have more than 28 incidences of verbal and physical violence.

$$P_{20} \text{ position} = \frac{20(10)}{100} = 2 \text{ round to position 2.5.}$$

$$\therefore P_{20} \text{ value} = \frac{13 + 14}{2} = 13.5$$

This means that 20% of the cartoons have less than 13.5 incidences.

Activity 4.16

The following data show the number of cars entering a parking area during a sample of ten-minute intervals.

10 22 9 24 27 29 9 23 12 31

Calculate Q_1 , Q_3 , P_{90} and P_{10} . Interpret your answers in context of the data.

Grouped data

Steps

1. Construct a frequency distribution with classes and frequencies.
2. Construct the cumulative 'less than' frequency column.
3. Determine the position of the quartile or percentile number.

$$\text{position } Q_j = \frac{jn}{4}$$

$$\text{position } P_j = \frac{jn}{100}$$

4. Compare the position of the fractile with the cumulative frequencies to determine which one of the intervals contains the fractile.

*image
not
available*

$$3. Q_j = L + \frac{\left(\frac{jn}{4} - F\right)c}{f_Q}$$

$$Q_1 = 3.5 + \frac{\left(\frac{1(255)}{4} - 32\right)3}{108} = 4.38$$

This means that 25% of the patients stay less than 4.38 days in hospital.

$$4. Q_3 = 6.5 + \frac{\left(\frac{3(255)}{4} - 140\right)3}{67} = 8.79$$

This means that 75% of the patients stay less than 8.79 days in hospital.

$$5. P_j = L + \frac{\left(\frac{jn}{100} - F\right)c}{f_P}$$

$$P_{85} = 9.5 + \frac{\left(\frac{85(255)}{100} - 207\right)3}{28} = 10.54$$

This means that 85% of the patients stay less than 10.54 days in the hospital.

Activity 4.17

The following frequency table shows the time (in minutes) taken to travel to work for a sample of 25 people from Gauteng. Calculate Q_1 , Q_3 , P_{90} and P_{10} . Interpret your answers in context of the data.

Class boundaries	f
15.5 – <21.5	2
21.5 – <27.5	6
27.5 – <33.5	8
33.5 – <39.5	4
39.5 – <45.5	4
45.5 – <51.5	1
Total	25

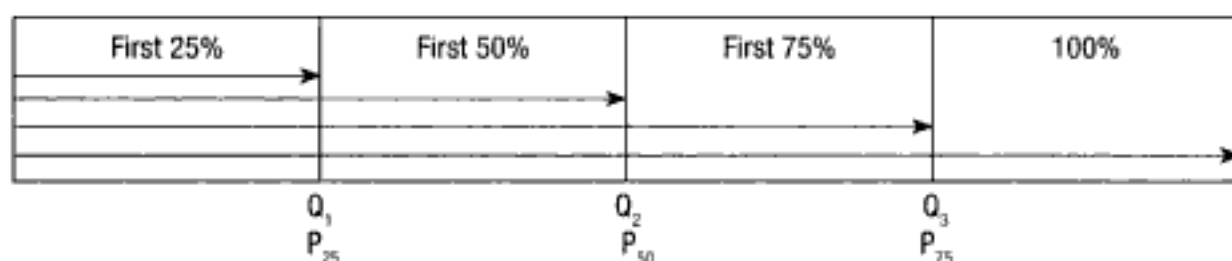
4.6 Measuring dispersion using measures of relative standing

These measures of dispersion are resistant to the effect of extreme numbers. They are frequently used in skewed distributions to give an estimate of spread.

4.6.1 Interfractile ranges

These ranges are measures of the spread between two fractiles in a distribution.

1. The interquartile range includes approximately the middle 50% of values and is the difference between the Q_3 and Q_1 values. This means that the first 25% and the last 25% of the data are cut off. Large values of this statistic indicate that the first and third quartiles are far apart, indicating a high level of variability.



The interquartile range = $Q_3 - Q_1$ ∴ the middle two quarters

2. A middle range is the middle proportion of the data between two percentiles with the cut-off portions at the beginning of the data set and the end of the data set, equal.
 - Middle 80% range = $P_{90} - P_{10}$
[This means that the first and the last 10% of the data are cut off]
 - Middle 40% range = $P_{70} - P_{30}$
[This means that the first and the last 30% of the data are cut off]

Example 4.21

Refer to example 4.19: a psychologist randomly selected 10 TV cartoon shows and counted the number of incidents of verbal and physical violence in each. The counts were as follows:

26 13 16 21 15 31 15 30 14 11

Q_1 value = 14

Q_3 value = 26

P_{80} value = 28

P_{20} value = 13.5

Interquartile range = $Q_3 - Q_1 = 26 - 14 = 12$

Middle 60% range = $P_{80} - P_{20} = 28 - 13.5 = 14.5$

Activity 4.18

The following data show the number of cars entering a parking area during a sample of ten-minute intervals. Calculate the middle 80% range, middle 70% range and the middle 60% range.

10 22 9 24 27 29 9 23 12 31

*image
not
available*

*image
not
available*

- It is compact, and provides information about centre, spread, symmetry and the presence of outliers.
- These plots are particularly useful when you want to compare several sets of related data.

Construction of a box-and-whisker plot

1. Draw a horizontal x -axis which covers the range of the data values.
2. Do the five-number summary table.
3. At any point on the y -axis draw a rectangular box horizontal to the x -axis whose left edge is at the lower quartile value (Q_1) and whose right edge is at the upper quartile value (Q_3). The box width is the interquartile range and shows the spread of the middle 50% of the data.
4. Draw a vertical line inside the box to indicate the position of the median.
5. Draw whiskers (horizontal lines) from the midpoints from each end of the box out to the smallest number and to the most extreme point – the minimum and maximum data values which is the range or the spread of the entire data set.
6. Place marks at distances 1.5 times the interquartile range from either end of the box – these are known as the inner fences.
7. Outliers are data values between the inner fences and the smallest and largest values.
8. A box plot also shows the symmetry or skewness of a distribution. In a symmetric distribution, the Q_1 and Q_3 values are equally distant from the median. If the distribution is skewed to the right the Q_3 will be farther away from the median than the Q_1 . If the distribution is skewed to the left, Q_3 will be closer to the median than Q_1 .

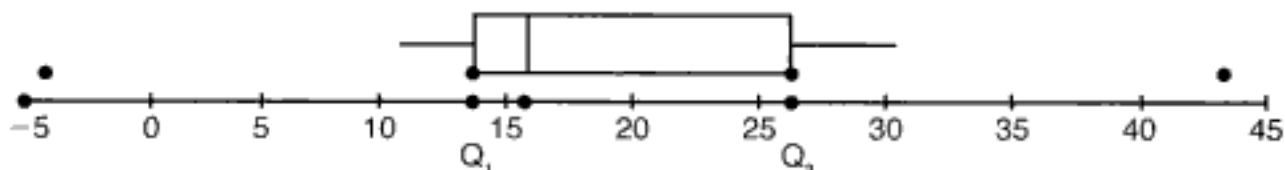
Example 4.24

A psychologist randomly selected 10 TV cartoon shows and counted the number of incidents of verbal and physical violence in each. The counts were as follows:

11 13 14 15 15 16 21 26 30 31

The five-number summary table:

1. The smallest data value $S = 11$
2. The lower quartile $Q_1 = 14$
3. The median $\text{Med} = 15.5$
4. The upper quartile $Q_3 = 26$
5. The largest data value $L = 31$
6. Left inner fence: $14 - (1.5 \times 12) = -4$
7. Right inner fence: $26 + (1.5 \times 12) = 44$



Interpretation

The five-number summary values are all indicated on the plot:

- Outliers are any observations larger than $26 + 1.5(16) = 54$ or smaller than $14 - 1.5(12) = 18$. The whisker to the left extends to 11, which is the smallest value and not an outlier. The whisker to the right extends to 31, which is the largest value and not an outlier. That means there are no outliers.
- The Q_1 is closer to the median than Q_3 , therefore the distribution is positively skewed.

Activity 4.21

The number of cars entering a parking area during a sample of ten-minute intervals is given below. Portray the five-number summary table for the distribution as a box plot.

10 22 9 24 27 29 9 23 12 31

TEST YOURSELF 4

For the distributions below, calculate and interpret in the context of the data (if possible):

- arithmetic mean
- median
- mode
- range
- mean absolute deviation
- standard deviation
- variance
- coefficient of variation
- Pearson's second coefficient of skewness
- draw graphs to determine median and mode
- interquartile range
- quartile deviation
- box plot.

1. The following data represent the pulse rates (beats per minute) of nine students enrolled for the statistics course:

76 60 60 81 72 80 80 68 73

2. In a research study concerning the long-term effectiveness of nicotine patches on participants who had previously smoked 20 cigarettes per day, a sample of 15 participants reported that they now smoke the following number of cigarettes per day:

10 10 7 10 10 9 8 10 9 8 6 9 8 10 8

3. The amount of aluminum contamination (in parts per million) in plastic wraps used to cover food, was recorded for a sample of 26 plastic specimens:

172 102 30 182 115 30 60 118 183
 63 119 191 70 119 222 79 120 244
 87 125 291 96 140 511 101 145

4. During a quality assurance check, the actual coffee content (in grams) of six jars of instant coffee was recorded as:

82.9 76.9 88.0 82.5 82.4 82.8

5. A sample of apples, guavas and mangos were analysed for the pesticide residues in the fruit. The amounts, in mg/kg of a certain pesticide were as follows:

0.2 1.6 4.0 5.4 5.7 11.4 0.2 3.4 2.4 6.6 4.2 2.7

6. During one month, records show the following results for the number of workers absent per day:

13 14 9 17 21 10 15 22 19 13 5
 22 13 19 23 17 21 10 9 20 18

7. The daily sales of a small business (in R'000) are given below for an 8 day period:

8.2 11.5 10.1 9.4 15.1 6.1 10.3 12.3

8. The following sample of lifetimes (in hours) of a certain type of battery used in a remote control is recorded as follows:

5.5 5.1 6.2 6.5 5.8 5.6 5.8 6.0

9. Corrosion of reinforcing steel is a serious problem in concrete structures located in the coastal areas. Researchers have been investigating the use of reinforced bars made of composite material. In one study glass-fibre-reinforced plastic bars were bonded to concrete. The following data were recorded on measured bond strength:

12.1 9.9 7.8 6.2 6.6 7.0 5.5 5.1 5.2 4.8 15.2 3.8
 5.4 5.2 4.9 10.7 13.1 8.5 3.4 20.6 13.8 12.6 4.1 8.9

10. The following frequency distribution summarises a sample of daily sales in (R'000):

Sales (R'000)	Frequency
96.5 – <96.9	1
96.9 – <97.3	8
97.3 – <97.7	14
97.7 – <98.1	22
98.1 – <98.5	19
98.5 – <98.9	32
98.9 – <99.3	6
99.3 – <99.7	4

11. The number of minutes after their appointment times each of a random sample of 64 patients had to wait to be served in a major local health facility were observed as follows:

Waiting time	Number of patients
0 – <4	10
4 – <8	17
8 – <12	16
12 – <16	14
16 – <20	7

12. The amount of caffeine in a sample of 250 ml servings of brewed coffee is summarised in the table below:

Caffeine (mg)	Number of cups
59.5 – <81.5	1
81.5 – <103.5	12
103.5 – <125.5	25
125.5 – <147.5	10
147.5 – <169.5	2

13. The lecturer in computer science recorded the amount of computer time (in minutes) needed by each student to complete an assignment:

Time	Number of students
0.1 – <0.5	3
0.5 – <0.9	10
0.9 – <1.3	16
1.3 – <1.7	9
1.7 – <2.1	5

14. A study of the number of trips on a particular day for a sample of 40 taxi drivers revealed the following data:

Number of trips	Frequency
0 – <5	3
5 – <10	6
10 – <15	8
15 – <20	13
20 – <25	7
25 – <30	3

15. A factory manager records the yearly sick leave (rounded to the nearest half day) taken by his employees:

Number of days	Number of employees
0 – <2.5	13
2.5 – <5.0	7
5.0 – <7.5	17
7.5 – <10.0	10
10.0 – <12.5	3
12.5 – <15.0	2

16. The engineering division of Continental Motors planned a campaign to improve plant safety. In preparation, the following accident data were compiled for a sample of 50 weeks:

Number of accidents	Number of weeks
0 – <5	6
5 – <10	25
10 – <15	11
14 – <20	7
20 – <25	1

17. The Strongbo Rubber Company has two factories. Both factories employ students during the holiday seasons. In factory A, the students are paid on average R982 per week with a standard deviation of R158. In factory B, the students earn on average of R1 208 per week with a standard deviation of R214. Which factory has the greatest relative dispersion?
18. Data have been collected on the life (in hours) of two brands of light bulbs. Compare the two brands using the coefficient of variation.

Brand A	Brand B
Mean = 5 800	Mean = 5 770
Standard deviation = 5 100	Standard deviation = 5 60

19. The operations manager of a package delivery service is deciding whether to purchase a new fleet of trucks. When packages are stored in the trucks in preparation for delivery, you need to consider two major constraints – weight and the volume for each item. He samples 200 packages and finds that the mean weight is 13.0 kg with a standard deviation of 1.7 kg. The mean volume is 0.8 cubic meter with a standard deviation of 0.2 cubic meter. Compare the variation of the weight and of the volume.

UNIT 5

Index numbers

Index numbers are used in business and economics as indicators to measure how much an economic variable changes over time or differs between two locations.

After completion of this unit you will be able to:

- calculate simple index numbers
- calculate a composite index number
- know the difference between unweighted and weighted index numbers
- change the base period
- calculate link relatives and percentage point changes
- understand the consumer price index.

An index number measures the change in a variable over time relative to the value of the variable during a pre-selected base period. The reason we are concerned with past changes is that we base business forecasts on what has happened in the past.

Index numbers are commonly used in business and economics as indicators of changing business or economic activity. Some examples of well known index numbers in South Africa are:

- Consumer price index (CPI)
- Producer price index (PPI)
- New car sales index
- JSE indexes
- Unemployment rate

The types of indexes that dominate business and economic applications are:

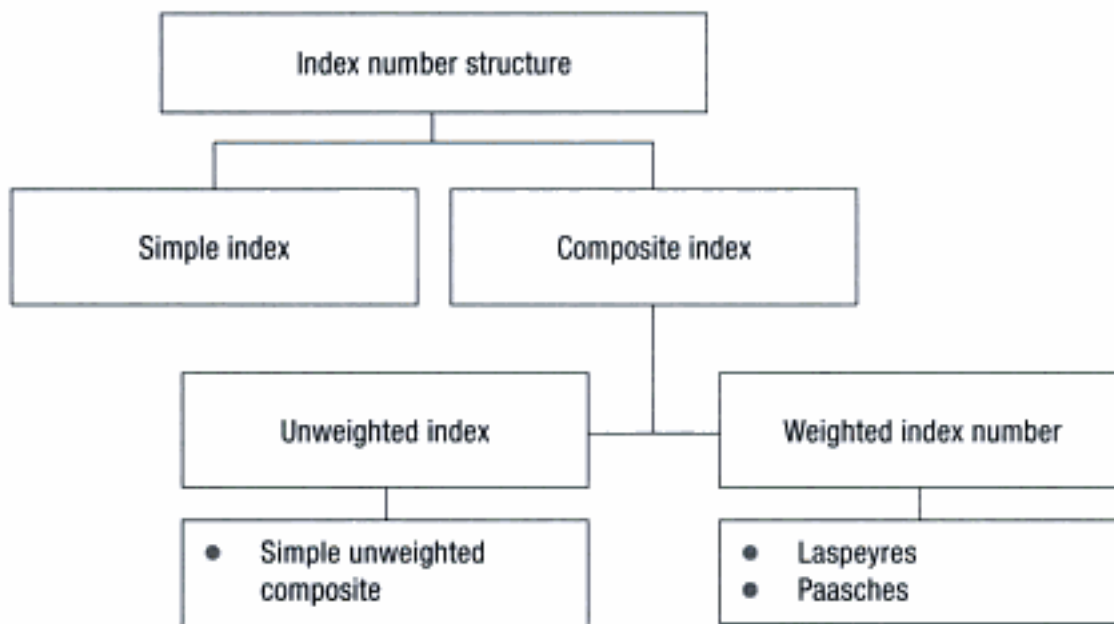
1. Price indexes (I_p) which are the most frequently used and measure changes in the price of a commodity or group of commodities between two time periods.
2. Quantity or volume indexes (I_q) which measure how much the quantity of a commodity or group of commodities changes over time.

In constructing an index number to describe the relative changes in commodity prices or quantities, we must first choose a base period. The base period is the time in the past against which all comparisons are made.

1. This period should be recent enough so that comparisons are not greatly affected by changing technology and consumer behaviour.

- This period should be one of relative economic stability without any abnormal influences.
- The base year index is always given the value of 100 and each subsequent year will have a value above or below 100 depending on whether there has been an increase or decrease in the data compared with the base year. For example, an index number of 124 will mean an increase of 24% since the base year and an index number of 93 will mean a decrease of 7% since the base year.

An index number that represents a comparison for an individual item is a **simple index number**. In contrast, when the index number has been constructed for a group of items, known as a basket of goods, it is an **aggregate** or **composite index number**.



5.1 Construction of a simple index number

A simple index number measures the changes (as a percentage) in price or quantity of a single item over time. It is calculated by dividing the current value (numerator) by a base value (denominator) and then multiplying the result by 100.

Steps

- Obtain the prices or quantities for the commodity over the time period of interest.
- Select the period to be used as base.
- Divide the current price (P_i) of the commodity by the base price (P_b).
- Multiply this ratio by 100
- Price index (I_p) = $\frac{\text{current year price}}{\text{base year price}} \times 100$

$$I_p = \frac{P_i}{P_b} \times 100$$

P_i represents the current period price and P_b the base period price.

6. The formula for a quantity index can be obtained by interchanging the values of P and Q in the price index formulae:

$$I_q = \frac{Q_t}{Q_b} \times 100$$

Q_t represents the current period quantity and Q_b the base period quantity.

Example 5.1

If milk costs R3.50 a liter in 2005 and R3.85 in 2006, the simple price index for 2006 will be

$$\begin{aligned} I_p &= \frac{P_t}{P_b} \times 100 \\ &= \frac{R3.85}{R3.50} \times 100 \\ &= 110 \end{aligned}$$

This means that milk increased in price by 10% over the period under consideration.

Activity 5.1

The following table provides the price per kilogram and the quantities purchased of nuts during the years 2007 and 2008. Use 2007 as base and construct simple price and quantity indexes per commodity for 2008.

	Prices per kg (rands)		Quantities purchased (kg)	
	2007	2008	2007	2008
Peanuts	45	50	600	550
Pecans	55	60	300	350
Cashews	60	70	325	325

5.2 Construction of composite (or aggregate) index numbers

These indexes are used to measure the relative change for a basket of related commodities.

- If each item in the overall index is of equal importance, the index is said to be **unweighted** – only prices *or* quantities are used in the construction of the index number. This particular index ignores both consumption patterns and the units to which prices refer.

- If each item in the overall index is not of equal importance, the index is said to be a **weighted** index. A system of weights is applied so the index will reflect the relative importance of its components. The price of an item is generally weighted by the quantity sold during that period. Prices *and* quantities are used in the construction of a weighted index.

5.2.1 Unweighted index numbers

Simple unweighted composite index number

All the commodities in the group are included in the calculation. This index considers each commodity in the basket as equally important, which results in the most expensive commodities, as well as the least consumed commodities, in the basket being overly influential.

$$I_p = \frac{\sum P_i}{\sum P_b} \times 100$$

Steps

1. Obtain the prices for the commodity over the time period of interest.
2. Select the period to be used as base.
3. Sum the prices of all the items in the given period (P_i).
4. Sum the prices of all the items in the base period (P_b).
5. Divide the numerator by the denominator.
6. Multiply the result by 100.
7. Interpret the answer.
8. If you want to calculate a simple unweighted quantity index, apply the following formula:

$$I_q = \frac{\sum Q_i}{\sum Q_b} \times 100$$

Example 5.2

The following table shows the costs of course material and price per unit that a student needs for a course in statistics:

	2007	2008
	P_b	P_i
Textbook	203	229
Calculator	80	70
Answer manual	10	10
Total	293	309

$$\begin{aligned}
 I_p &= \frac{\sum P_t}{\sum P_b} \times 100 \\
 &= \frac{309}{293} \times 100 \\
 &= 105.46
 \end{aligned}$$

This means that the prices increased by 5.46% over the period under consideration.

Activity 5.2

Construct a simple unweighted price and quantity index for 2008.

	Prices per kg (R)		Quantities purchased (kg)	
	2007	2008	2007	2008
Peanuts	45	50	600	550
Pecans	55	60	300	350
Cashews	60	70	325	325

5.2.2 Weighted composite index numbers

These methods take into account the different consumption (or quantity) levels of the commodities in the basket of goods. An important question in applying the process is: which quantities are used?

Laspeyres and Paashes index numbers are the most popular composite index numbers.

Laspeyres index

This method uses quantities consumed during the base period as a weighting factor and assumes that whatever the price changes, the quantities purchased will remain the same. Using this method will be misleading when the purchase quantities change significantly from those in the base period. One solution to the problem of purchase quantities that change relative to those of the base period is to change the base period regularly, so that the quantities are regularly updated.

The best-known Laspeyres index is the consumer price index (CPI).

$$I_p(L) = \frac{\sum P_t Q_b}{\sum P_b Q_b} \times 100$$

Steps

1. Collect price and quantity information for each of the items to be used in the composite index.
2. Select the base period.
3. Denote the current prices and quantities as P_i and Q_i respectively.
4. Denote the base prices and quantities as P_b and Q_b respectively.
5. Multiply the current period price (P_i) by the base period quantity (Q_b) for each item and sum the resulting values ($\sum P_i Q_b$).
6. Multiply the base period price (P_b) by the base period quantity (Q_b) for each item and sum the resulting values ($\sum P_b Q_b$).
7. Divide the first sum by the second sum and multiply the result by 100.
8. Interpret the answer.

Paashes' index

This method uses quantities consumed during the current period as a weighting factor. It measures the change in total cost of goods that represent a consumption pattern typical of the current year, and therefore avoids the problem of changing consumption patterns.

$$I_p(P) = \frac{\sum P_i Q_i}{\sum P_b Q_i} \times 100$$

Steps

1. Collect price and quantity information for each of the items to be used in the composite index.
2. Select the base period.
3. Denote the current prices and quantities as P_i and Q_i respectively.
4. Denote the base prices and quantities as P_b and Q_b respectively.
5. Multiply the current period price (P_i) by the current period quantity (Q_i) for each item and sum the resulting values ($\sum P_i Q_i$).
6. Multiply the base period price (P_b) by the current period quantity (Q_i) for each item and sum the resulting values ($\sum P_b Q_i$).
7. Divide the first sum by the second sum and multiply the result by 100.
8. Interpret the answer.

Example 5.3

The following table shows a farmer's orchard production. Construct Laspeyres' and Paashes' price indexes for the year 2008 with 2007 as the base.

	Price per box		Boxes ('000)		P_0Q_0	P_0Q_1	P_1Q_0	P_1Q_1
	2007 P_0	2008 P_1	2007 Q_0	2008 Q_1				
Apples	20	25	2.0	1.5	50.0	40.0	30.0	37.5
Oranges	15	17	1.4	1.6	23.8	21.0	24.0	27.2
Mangos	40	35	5.0	2.0	175.0	200.0	80.0	70.0
Bananas	30	38	7.0	9.0	266.0	210.0	270.0	342.0
					514.8	471.0	404.0	476.7

$$\begin{aligned}
 I_p(\text{Laspeyres}) &= \frac{\sum P_1 Q_0}{\sum P_0 Q_0} \times 100 \\
 &= \frac{514.8}{471.0} \times 100 \\
 &= 109.30
 \end{aligned}$$

Price increase of 9.30% since 2007.

$$\begin{aligned}
 I_p(\text{Paashes}) &= \frac{\sum P_1 Q_1}{\sum P_0 Q_1} \times 100 \\
 &= \frac{476.7}{404.0} \times 100 \\
 &= 118
 \end{aligned}$$

Price increase of 18% since 2007.

Activity 5.3

Construct weighted indexes for 2008.

	Prices per kg (R)		Quantities(kg)	
	2007	2008	2007	2008
Peanuts	45	50	600	550
Pecans	55	60	300	350
Cashews	60	70	325	325

5.3 Additional topics on index numbers

5.3.1 The consumer price index (CPI)

This index is a very important economic indicator and is used to determine inflation rate and cost of living. Its monthly publication by the Department of Statistics is a matter of great publicity.

The formula used in determining the CPI in South Africa is that of Laspeyres.

$$\text{CPI} = \frac{\sum P_t Q_b}{\sum P_b Q_b} \times 100$$

To determine the base year weight factor, at least 10 000 households were sampled out of different income groups and metropolitan areas. Following international practice, the base period used for the CPI must change at least every five years. A monthly index for each consumer item for each area is determined by making use of the above formula and then a combined CPI is calculated.

The formula for determining inflation rate is as follows:

$$\text{inflation rate} = \frac{\text{CPI current year}}{\text{CPI of corresponding month of previous year}}$$

5.3.2 Percentage points change

One convenience of index numbers is that by starting from 100, the percentage change from the base period is found just by subtraction. The differences thereafter are referred to as percentage points.

Example 5.4

The percentage increase is

Year	Index	Percentage points increase	Percentage increase
1996	100		
1997	120	20	20%
1999	160	40	33.33%
2000	188	28	17.5%

- $\frac{120 - 100}{100} \times 100 = 20\%$
- $\frac{160 - 120}{120} \times 100 = 33.33\%$

As the index gets larger the same percentage change is represented by a larger difference. A change from 100 to 120 is the same as a change from 300 to 360, but the impression can be very different. In practice the solution will be to change the base year.

5.3.3 Link relatives

Link or chain relatives are indexes for which the base is always the preceding period. Therefore, each index number represents a percentage comparison with the preceding year. Such indexes are useful for showing year to year comparisons, but not as a basis for making long-run comparisons.

Example 5.5

The rand value of sales for the Papillion Café between 2002 and 2007 is as follows:

Month	January	February	March	April	May
Sales (R)	14 980	16 433	20 194	23 015	23 621
Link relative		109.7	122.9	114.0	102.6

- February link relative = $\frac{16\,433}{14\,980} \times 100 = 109.7$
- March link relative = $\frac{20\,194}{16\,433} \times 100 = 122.9$

Activity 5.4

The following table lists the retail price of milk per litre between June and October at a local grocery store. Determine the link relatives for milk.

June	July	August	September	October
1.99	2.50	2.50	2.99	3.40

5.3.4 Changing the base year

Changing the base period is necessary under a number of conditions:

- if the original base period is too long ago
- if the two indexes you want to compare have different base periods
- inclusion of new items in the index or disappearance of old ones
- new techniques
- abnormal influences on the base period.

The base period for index numbers can be shifted by dividing each original index by the index of the newly designated base year and multiplying the result by 100.

$$\text{new index number} = \frac{\text{original (old) index number}}{\text{old index for new base}} \times 100$$

Example 5.6

Value indexes for the Ford Motor Co.

	2000	2001	2002	2003	2004	2005
2002 = 100	74.2	81.4	100.0	114.0	117.0	118.9
2004 = 100	63.4	69.6	85.5	97.4	100.0	101.6

$$2000 = \frac{74.2}{117.0} \times 100 = 63.4$$

$$2001 = \frac{81.4}{117.0} \times 100 = 69.6$$

$$2002 = \frac{100.0}{117.0} \times 100 = 85.5$$

Activity 5.5

For the following consumer price indexes, move the base to 2004.

2000	2001	2002	2003	2004	2005	2006
100	104.2	109.8	116.3	121.3	120.0	117.4

TEST YOURSELF 5

Calculate all simple and composite index numbers for questions 1 to 5.

1. A lecturer's essential commodities for teaching statistics:

	Unit	Unit prices		Quantities	
		2007	2008	2007	2008
Chalk	box	5.00	5.50	4	5
Red pen	per 1	0.72	0.75	3	5
Text book	1	103.00	139.00	1	1
Aspirin	bottle	12.00	10.00	3	4

2. The Valdo Art School has compiled the following information showing prices and quantities of the following supplies for 2007 and 2008:

	Prices		Quantities	
	2007	2008	2007	2008
Brushes	4.74	4.92	50	43
Canvas	3.10	3.41	920	907
Oil	6.29	6.83	107	121

3. The table below shows the prices and annual consumption of the raw materials used in Gauteng Breweries in 2007 and 2008:

	Prices		Unit Quantities	
	2007	2008	2007	2008
Malt	49	46	10874	15116
Hops	512	724	732	696
Sugar	46	51	1865	2486
Wheat flour	31	27	873	1093

4. Tixif Limited sells three types of chain saws. Company records showed the prices (R'000) and quantities sold as follows:

	Price (R)		Quantity	
	2007	2008	2007	2008
X	30	40	22	30
Y	50	60	31	40
Z	120	99	8	12

5. Mr Hiram, a pensioner, has kept a record on the costs of certain items purchased weekly:

	Price per unit		Quantities Purchased	
	2007	2008	2007	2008
Coffee	12.00	15.00	30	32
Cookies	5.60	4.99	7	9
Sugar	7.50	8.20	20	24

6. Mr Rolling has been offered a job in Cape Town with a salary of R123 500 per year. The cost of living index is 132. If he presently earns R100 000 per year in Johannesburg with a cost of living index of 120, will he be financially better off in the new job?
7. The CPI values for the first eight months of 2008 with 2006 = 100 were:

233 236 240 243 248 249 252 255

Shift the base of the index to June 2007 and determine the purchasing power of the rand per month for both index series.

8. The producer price index with 2006 = 100 for the previous twelve months were:

181 201 221 227 234 238 245 249 260 268 290 300

Calculate the percentage point increases and the percentage increases in the index numbers.

9. The reported new cases of tuberculosis in a busy hospital were as follows:

January	February	March	April	May	June
239	311	289	264	321	199

Calculate the link relatives and then the percentage point and percentage increases.

10. The following figures relate to the library expenditure (R) of a small town. Also given is the retail price index per year:

Year	2000	2001	2002	2003
Expenditure	4 800	5 230	5 800	6 700
Retail index	103	110	120	128

If the retail index is taken into consideration, was there a real increase in the library expenditure?

UNIT 6

Summarising bivariate data: regression and correlation analysis

This unit deals with methods to summarise data consisting of observations on two quantitative variables (bivariate data) presented as ordered pairs. The purpose is to understand if there is a relationship between two variables and what you can do if a relationship exists.

After completion of this unit you will be able to:

- explain the purpose of regression and correlation analysis
- compute and explain the meaning of the coefficients of correlation and determination
- compute the regression equation and use this measure to do estimates
- compute correlation by making use of ranking.

Regression and correlation analysis are statistical tools used to study the relationship between two variables of which one is dependent and the other independent. It is used to determine:

- whether there is a relationship between the variables
To describe the relationship between the two variables, we first graphically represent the data in a scatter diagram. This visual representation can give an immediate impression of a set of data; it will illustrate whether there is a relationship and also suggest whether the relationship is linear, positive or negative. The strength of the relationship may be concluded tentatively.
- how good that relationship is
The correlation coefficient is a numerical descriptor of the data that is used to measure how good this relationship is.
- how the relationship can be used to make predictions.
Once the scatter diagram and correlation coefficient indicate that a linear relationship exists between the two variables, we proceed to find a linear equation that describes the relationship between the two variables. This equation can be used to make predictions within the given range of the data (interpolation).

Note: This unit will deal with linear relationships involving two variables only.

6.1 Response variable (y) and explanatory variable (x)

Each observation of bivariate data can be thought of as a data pair of the form (x, y) . The x is the explanatory (or independent) variable and y is the response (or dependent) variable. The response variable y is the variable whose value depends on, or can be explained by, the value of the explanatory or independent x -variable. When we analyse data on two variables, the first step is to distinguish between the response variable and the explanatory variable.

Activity 6.1

Identify the independent x - and dependent y -variables in each of the following:

1. size of advertisements of companies in the yellow pages and the number of calls to the business that were generated by the advertisement
2. number of hours of training per employee and revenue of the company
3. hours of training to use Excel and number of errors made using the program
4. temperature and ice-cream sales
5. weight and calorie intake
6. circulation of a magazine and advertising charges
7. the hardness and tensile strength of die-cast aluminium
8. wattage of a heater and the effective heating area
9. number of cavities and sugar intake of children
10. number of police on the streets and number of crimes.

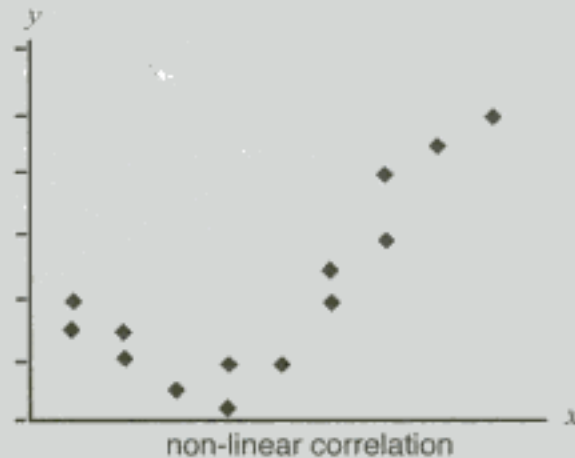
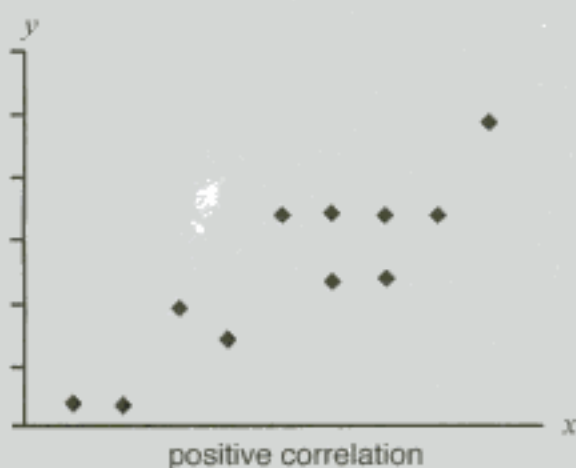
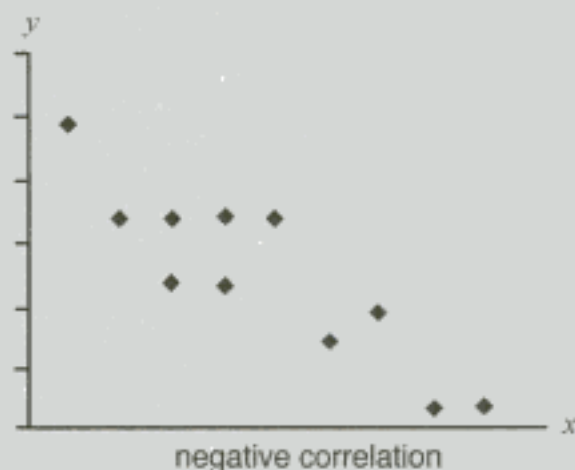
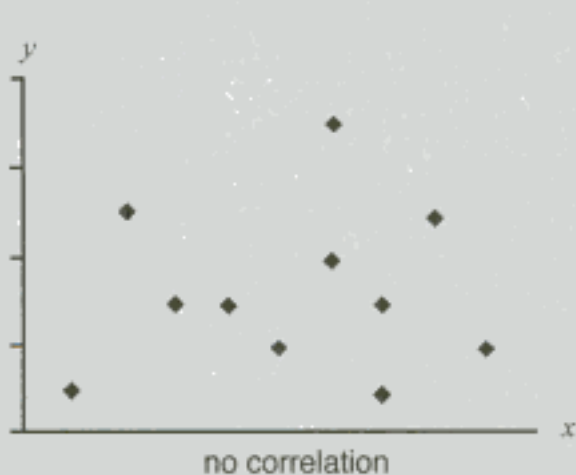
6.2 Scatter diagram

A scatter diagram (or scatter plot) is a graph of data from two quantitative variables (bivariate data). This graph can identify the type of relationship that might exist between two variables.

Steps

1. Collect pairs of data (x, y) . The data are paired in a way that matches each value from one data set with a corresponding value from a second data set.
2. Select which variable is the dependent (y) variable and which is the independent (x) variable. The label y goes to the variable which we want to predict. The other variable is then labeled as x .
3. Arrange the data in two columns, x and y .
4. Draw a set of axes.
5. The horizontal axis represents the x -variable and is scaled so that any x value can be easily located.
6. The vertical axis represents the y -variable and is scaled so that any y value can be easily located.

7. Each pair of observations (x, y) is plotted as a point. That is where a vertical line from the value on the x -axis meets a horizontal line from the value on the y -axis.
8. The points are not connected.
9. Scatter plots can take on the following patterns:
 - The plot can show no relationship, because no pattern can be identified.
 - The plot can show a positive relationship because the dots start at the bottom left and move upwards to the top right. Although the data points do not fall exactly on a line, they appear to cluster about a line. A positive relationship means that if the x -variable increases, the y -variable will also increase.
 - The plot can show a negative relationship because the dots start at the top left and move downwards to the bottom right. A negative relationship means that if the x -variable increases, the y -variable will decrease.
 - If all the points fall exactly along a straight line, in a negative or positive direction, we say the relationship is perfect.
 - Non-linear relationships are beyond the scope of this unit.



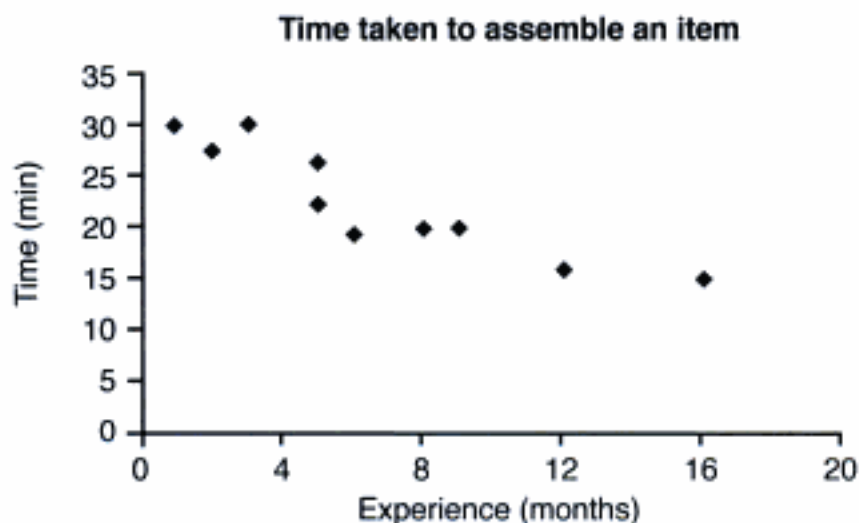
Example 6.1

You timed how long it takes ten workers to assemble an item. It was possible for you to match these times with the length of the workers' experience (in months). The results obtained are shown below:

Person	Experience (months) x	Time (min) y
A	2	27
B	5	26
C	3	30
D	8	20
E	5	22
F	9	20
G	12	16
H	16	15
I	1	30
J	6	19
	67	225

Scatter plot for the data

The time taken to assemble the item is the y -variable because time depends on the experience (x) of the worker.



The scatter plot shows a negative relationship, which means the more experienced workers take a shorter time to assemble the product.

Activity 6.2

During the baking of a certain type of bread roll on very low heat, each bread roll goes through a series of heat processes. The length of time spent under this heat treatment is related to the lifespan of the bread rolls. A sample of eight bread rolls that underwent different baking times were selected and the life span (in hours) of each was recorded:

Length of time	18	13	18	15	10	12	8	4
Life span	23	20	18	16	14	11	10	7

Draw a scatter diagram and interpret the relationship in the context of the given data.

6.3 Correlation analysis (r)

A correlation exists between two variables when one of them is related or can be influenced by the other in some way.

The linear correlation coefficient is a numerical measure to describe the degree of strength and direction by which one variable is related to another. A linear relationship means that when graphed, the points approximate a straight line pattern.

We use an equation, known as Pearson's correlation coefficient, to measure this strength. This correlation coefficient is represented by the small letter r .

Steps

1. List the values of the x - and the y -variables in two columns.
2. Sum the x values (Σx) and sum the y values (Σy).
3. Square each of the x values and add the column (Σx^2).
4. Square each of the y values and add the column (Σy^2).
5. Multiply each x with its corresponding y value and add the products (Σxy).
6. Substitute these values into the formula for r and determine the value of r .

$$r = \frac{n \cdot \Sigma xy - \Sigma x \Sigma y}{\sqrt{[n \cdot \Sigma x^2 - (\Sigma x)^2][n \cdot \Sigma y^2 - (\Sigma y)^2]}}$$

6.3.1 Characteristics of the linear correlation coefficient

1. When the slope of the scattered points is negative, the r -value is negative and when it is positive, the r -value is positive. Thus, the sign of r indicates the direction of the relationship between the variables x and y .
2. A correlation coefficient can range in value from -1 to $+1$. The closer to $+1$ or -1 , the better the relationship. If r is close to 0 , there is little or no linear correlation between the two variables.
3. The strength of the correlation is not dependent on direction. $r = 0.95$ and $r = -0.95$ are equal in strength. The absolute value of the coefficient reflects the strength of the correlation; a correlation of -0.7 is stronger than a correlation of $+0.3$.
4. A correlation coefficient is a measure of association without a unit.

6.3.2 Interpreting a correlation coefficient (r)

- If an inverse (negative) relationship exists, that is, if y decreases as x increases, then r will fall between 0 and -1 .
- If there is a direct (positive) relationship, that is, if y increases as x increases, then r will fall between 0 and $+1$.
- If there is no relationship between x and y , then $r = 0$.
- This measure enables us to make statements such as: the correlation is strong, weak, etc.

Size of r	General interpretation
$\pm(0.9 \text{ to } 1.0)$	Very strong relationship
$\pm(0.8 \text{ to } 0.9)$	Strong relationship
$\pm(0.6 \text{ to } 0.8)$	Moderate relationship
$\pm(0.2 \text{ to } 0.6)$	Weak relationship
$\pm(0.0 \text{ to } 0.2)$	Very weak or no relationship

- If the association between two variables is strong, then knowing the one variable helps a lot in predicting the other. But when there is a weak association, information about one variable does not help much in estimating the other.

6.3.3 The coefficient of determination (r^2).

This is a more conservative measure of the relationship because it enables us to calculate the total variation in the y -variable that is explained by the corresponding variation in the x -variable. To determine the value of this measure, simply square the correlation coefficient and multiply the answer by 100.

This answer will always be positive within the range 0 to 100%.

$$\text{coefficient of determination} = r^2 \cdot 100$$

Example 6.2

You timed how long it takes ten workers to assemble an item. It was possible for you to match these times with the length of the workers' experience. The results obtained are shown below:

Person	Experience x	Time (min) y	xy	x^2	y^2
A	2	27	54	4	729
B	5	26	130	25	676
C	3	30	90	9	900
D	8	20	160	64	400
E	5	22	110	25	484
F	9	20	180	81	400
G	12	16	192	144	256
H	16	15	240	256	225
I	1	30	30	1	900
J	6	19	114	36	361
	67	225	1 300	645	5 331

1. Correlation coefficient:

$$r = \frac{10(1\,300) - (67)(225)}{\sqrt{[10(645) - (67)^2][10(5\,331) - (225)^2]}}$$

$$= -0.90$$

This is a good negative correlation.

2. Coefficient of determination:

$$r^2 \cdot 100 = (-0.90)^2 \cdot 100$$

$$= 81\%$$

81% of the total variation in the time taken to produce an item can be explained by the variation in experience. The remaining 19% is explained by other factors.

Activity 6.3

During the baking of a certain type of bread roll on very low heat, each bread roll goes through a series of heat processes. The length of time spent under this heat treatment is related to the lifespan of the bread rolls. A sample of eight bread rolls that underwent different baking times was selected and the life span (in hours) of each roll was recorded.

Length of time	18	13	18	15	10	12	8	4
Life span	23	20	18	16	14	11	10	7

Calculate the correlation coefficient and the coefficient of determination. Interpret your answer in the context of the given data.

6.4 Regression analysis

Once the scatter diagram and correlation coefficient indicates that a linear relation exists between the two variables, the next step is to determine the equation of the straight line that best describes the pattern of the relationship between the two variables. This equation, known as the regression equation, can be used to predict a dependent variable (y) if the independent variable (x) is known.

'**Best**' refers to how close the predictions of y (\hat{y}) are to the actual values of y .

A **linear** relationship between the two variables means that the equation will result in a straight line if plotted on a graph.

6.4.1 Formulating the regression equation

Any linear function involving two variables can be expressed in the form:

$$\hat{y} = a + bx$$

Where:

\hat{y} = estimated y value for a given x -value

a = intercept on the y -axis

b = the slope (the average change in y for each change of 1 unit in x)

Steps

1. Obtain a random sample of n data pairs (x, y) , with x the independent variable and y the dependent variable.
2. Use the data pairs to calculate n , Σx , Σy , Σxy and Σx^2 .
3. Calculate a and b values in the equation by making use of the method of least squares: the least squares principle states that the method used to determine the regression equation must be such that the sum of the squares of the differences between each actual y -value and the corresponding estimated y value is a minimum.
4. The ' b ' value is the slope of the straight line equation. The slope is the amount by which y increases or decreases when x increases by 1 unit.

$$b = \frac{n \cdot \Sigma xy - \Sigma x \Sigma y}{n \cdot \Sigma x^2 - (\Sigma x)^2}$$

5. The ' a ' value is the y -intercept; that is where $x = 0$ and the straight line crosses the y -axis.

$$a = \frac{\Sigma y}{n} - b \frac{\Sigma x}{n}$$

6. Substitute the a and the b values into the regression line equation: $\hat{y} = a + bx$

6.4.2 Using the regression equation to make predictions

- The regression equation $\hat{y} = a + bx$ allows you to use the independent variable x to make predictions for the dependent variable y .
- Substitute the independent x -value you require a prediction for into the regression equation and calculate the estimated y -value (\hat{y}).

6.4.3 Plot the regression line on the scatter diagram

Steps

1. For predictions using a graph, a straight line is fitted to the data. The best fitting straight line can be obtained by making use of the equation: $\hat{y} = a + bx$
2. Since we are dealing with a linear function, we only need to estimate two points, the rest of the \hat{y} -values will all fall on that straight line.
3. Choose any two x values from the x -scale on the scatter diagram. Substitute the two chosen values for x into the regression equation and obtain the two corresponding values for \hat{y} .
4. Plot the two coordinate points on the same axis as the scatter diagram.
5. Use a ruler and draw a straight line through the two points.
6. When the estimated \hat{y} is plotted against x and the points are connected, the result is called the regression line or the line of best fit.
7. To do a prediction for a specific x -value, find the required x on the x -axis; draw a vertical line to the regression line. From this point draw a horizontal line to the y -axis and read the estimate from the y -axis.

Note: Estimated y -values are meaningful only for x -values in (or close to) the range of the given data.

Example 6.3

You timed (in minutes) how long it takes ten workers to assemble an item. It was possible for you to match these times with the length of the workers' experience (months). The results obtained are shown below. Develop the regression equation and estimate the assembly time for a worker with four months experience and for a worker with 10 months experience.

Person	Experience x	Time y	xy	x^2	y^2
A	2	27	54	4	729
B	5	26	130	25	676
C	3	30	90	9	900
D	8	20	160	64	400
E	5	22	110	25	484
F	9	20	180	81	400
G	12	16	192	144	256
H	16	15	240	256	225
I	1	30	30	1	900
J	6	19	114	36	361
	67	225	1 300	645	5 331

1. Regression equation:

$$b = \frac{n \cdot \sum xy - \sum x \sum y}{n \cdot \sum x^2 - (\sum x)^2}$$

$$b = \frac{10(1\,300) - (67)(225)}{10(645) - (67)^2}$$

$$b = -1.06$$

$$a = \frac{\sum y}{n} - b \frac{\sum x}{n}$$

$$a = \frac{225}{10} - (-1.06)\left(\frac{67}{10}\right) = 29.60$$

$$\hat{y} = a + bx$$

$$\hat{y} = 29.60 + (-1.06)x$$

2. Interpret a and b :

The a value in the equation represent the y -intercept. That is the point on the y -axis where $x=0$. In this equation it means that a worker with no experience will take 29.6 minutes to assemble the product.

The b -value represents the slope, which means that for every additional month of experience, a worker will take 1.06 minutes less to assemble the product.

3. Estimates:

For a worker with four months experience:

$$\hat{y} = 29.60 + (-1.06)x$$

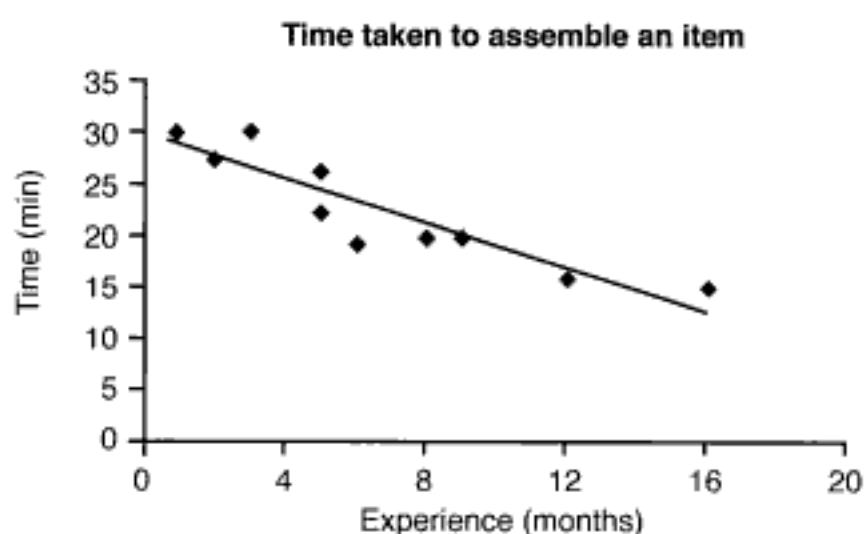
$$\hat{y}_{(x=4)} = 29.60 - 1.06(4) = 25.36 \text{ minutes}$$

For a worker with 10 months experience:

$$\hat{y}_{(x=10)} = 29.60 - 1.06(10) = 19 \text{ minutes}$$

4. Place the regression line on the scatter plot.

Plot the coordinates (4 ; 25.36) and (10 ; 19) obtained from the estimates on the scatter diagram and join the two points with a straight line.



Activity 6.4

During the baking of a certain type of bread roll on very low heat, each bread roll goes through a series of heat processes. The length of time spent under this heat treatment is related to the lifespan of the bread rolls. A sample of eight bread rolls that underwent different baking times was selected and the life span (in hours) of each roll was recorded.

Length of time	18	13	18	15	10	12	8	4
Life span	23	20	18	16	14	11	10	7

If a bread roll will spend 16 hours under heat treatment, how long do you expect it will remain fresh? Do your estimate using the regression equation and the regression line. Give the meaning of the slope.

6.5 Spearman rank correlation coefficient (r_s)

This coefficient measures the strength of the relationship between two variables on the basis of their ranks instead of values (ordinal data) and can be used in problems where one or both variables can be ranked even though they cannot be measured on a numerical scale. It can be interpreted in a manner similar to the correlation coefficient r but because a great deal of the data gets lost, it provides a less reliable result.

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

Steps

1. Rank the x and the y variables: assign numbers from 1 onwards to the data values, starting with the smallest (or largest) value up to the largest (or smallest) value. Keep each x together with its y . Remember to use the same type of ranking for x and y – from high to low or from low to high. If two values are the same, they are first assigned ranks (say 2 and 3) and then the average of the ranks is determined (2.5). That average is then assigned to each appropriate value.
2. Calculate the difference between ranks of the two variables (d).
3. Square these differences (d^2) and add the column.
4. Substitute the required values into the formula and calculate the rank-order correlation coefficient.
5. Interpret the coefficient.

Example 6.4

The safety officer of a company wants to know if experience influences the quality of an employee's work. She selects 10 employees at random and records their years of work experience and their quality rating as assessed by their supervisors.

Quality rating: 7 = excellent and 1 = poor

Assume that the employees with the longest years of experience will be the most highly rated. To keep the type of ranking the same, the shortest years of experience will be ranked the lowest.

x	y	Rating		d	d^2
Employee	Experience	x -code	y -code		
1	1	2	1	1	1
2	17	5	8	-3	9
3	20	5	9	-4	16
4	9	6	4.5	1.5	2.25
5	2	3	2	1	1
6	13	5	7	-2	4
7	9	4	4.5	-0.5	0.25
8	23	6	10	-4	16
9	7	3	3	0	0
10	10	6	6	0	0
					49.5

$$r_s = 1 - \frac{6(49.5)}{10(10^2 - 1)} = 0.7$$

The correlation is moderate and positive. This means that the more experience the employees have, the better their rating.

Activity 6.5

During the baking of a certain type of bread roll, advertised as having a long shelf-life, each bread roll goes through a series of heat processes. The length of time spent under this baking process is related to the shelf-life of the bread rolls. A sample of eight bread rolls that underwent different baking times was selected, and the shelf-life (in hours) of each roll was recorded.

Length of time	18	13	18	15	10	12	8	4
Shelf-life	23	20	18	16	14	11	10	7

Determine Spearman's rank correlation coefficient and interpret your answer.

TEST YOURSELF 6

For the data sets in questions 1 to 9:

- identify the dependent and independent variables
- determine if there is a relationship between the two variables using a scatter plot
- measure the strength of the relationship using Pearson's correlation coefficient
- determine the coefficient of determination and interpret its meaning in the problem
- assume a linear relationship and do the required estimate using the regression equation
- plot the line-of-best-fit on the scatter plot
- interpret the meaning of the y -intercept and the slope.

1. Potato chip lovers do not like soggy chips, so it is important to find characteristics of the production process that produce chips with an appealing texture. The following sample data on frying time (in seconds) and moisture content (%) were selected:

Frying time	65	50	35	30	20	15	10	5
Moisture content	1.4	1.9	3.0	3.4	4.2	8.1	9.7	16.3

Predict the moisture content of the chips after 40 seconds frying time.

2. From the following data, determine the resting pulse rate that you would expect from someone exercising for a daily average of (a) 45 minutes (b) 15 minutes and (c) 2.5 hours:

Daily exercise (min)	20	30	60	10	100	0	120	160	160	180
Pulse/min.	75	70	70	85	50	90	60	52	48	64

3. The following sample data measure levels of anxiety before a test and test marks obtained for the test.

Anxiety	23	14	14	0	17	20	20	15	21
Test %	43	59	48	77	50	52	46	51	51

What test marks do you expect if the anxiety level was 12?

4. A chemistry lab testing food has seven divisions that do different chemical tests on food products. The number of hours devoted to safety training and the number of hours lost due to industry-related accidents were recorded for each division:

Safety training	10	19	30	45	50	65	80
Hours lost due to accidents	80	65	68	55	35	10	12

After 60 hours of training, how many hours do you expect to lose due to accidents?

5. The Rip-Off Vending Machine Company operates coffee vending machines in several office buildings. The company wants to study the relationship that exists between the number of cups of coffee sold per day and the number of persons working in each office. Data for this study were collected by the company and are presented below:

Number of cups sold	10	20	30	40	30	20	40	40	50	10	40	20
Number of persons	5	6	14	19	15	11	18	22	26	4	23	10

Predict the number of cups of coffee that you will expect to sell if there are 45 people working in an office.

6. The following data reflect the family income and food expenditure (in R'000) of a sample of 10 low income families:

Income	24	15	18	12.6	8	9.5	21	11.4	6.4	13.2
Expenditure	3.6	2.9	2.9	2.6	2	2.4	3	2.5	1.8	2.4

Compute all relevant statistics by making use of the step-by-step procedure and estimate food expenditure of a family with an income of R20 000.

7. When buying almost any item, it is often advantageous to buy it in as large a quantity as possible. The unit price is usually less for the larger quantities. The data shown in the table were obtained to test this theory:

Number of units	1	3	5	10	15
Cost per unit	55	52	48	32	25

Estimate the price if you buy 14 items.

8. During the making of certain electrical components each item goes through a series of heat processes. The length of time (minutes) spent in this heat treatment is related to the useful life (hours) of the components. To find the nature of this relationship a sample of 10 components was selected from the process and tested to destruction:

Time in process	25	27	25	26	31	30	32	29	30	44
Length of life	2 005	2 157	2 347	2 239	2 889	2 942	3 048	3 002	2 943	3 844

Predict the useful life of a component that spends 33 minutes in process.

9. The following data show the present maintenance costs and age of a sample of eight similar machines used in a clothing factory:

Age	1	3	4	4	6	7	7	8
Maintenance (R)	200	550	650	800	1 150	1 100	1 300	1 500

Compute all relevant statistics by making use of the step-by-step procedure and estimate the maintenance cost of a five-year-old machine.

10. Below are the rankings of the top 10 products produced by Peter's Party Products for last year and this year:

Product	Last year	This year
Crackers	1	3
Hats	3	1
Masks	4	2
Balloons	6	10
Whistles	7	9
Streamers	8	7
Flags	9	8
Face paint	2	4
Joke food	5	6
Joke cards	10	5

Compute the Spearman correlation coefficient and interpret your answer.

11. Ten sales agents of a company had the following number of years of service:

Agent	A	B	C	D	E	F	G	H	I	J
Years	8	6	4	12	5	3	1	14	9	10

The manager of the company arranged the agents in the following order, from most excellent (H) to least excellent (F).

H	I	D	J	A	E	C	B	G	F
---	---	---	---	---	---	---	---	---	---

Determine the Spearman rank correlation coefficient between years of service and excellence.

UNIT 7

Time series

This unit discusses the general use of forecasting in business and several methods that are available for making forecasts.

After completion of this unit you will be able to:

- explain the purpose of time series analysis
- explain the components of the multiplicative model
- use linear models to analyse and project the trend of a time series
- measure the seasonal effect in a time series
- use time series in forecasting.

Forecasting is the science of predicting the future. It is used in the decision making process to help business people reach conclusions about buying, selling, producing and many other actions.

Time-series data are numerical data gathered on a given characteristic over a period of time at regular intervals. The objective is to analyse how observed data change over time, in order to detect patterns that will enable us to predict future values. Thus, time series analysis helps us cope with uncertainty about the future.

7.1 Components of a time series

It is generally believed that the factors that have influenced data in the past and present will continue to do so in more or less the same way in the future. The major goal of time-series analysis is to isolate and measure these influencing factors for forecasting purposes.

Four separate components – trend, cyclical, seasonal and irregular – are combined to provide specific values for the time series.

- Secular trend (T) is the underlying long-term movement (increase or decrease) over time in the value of the data recorded and is usually the result of long-term factors such as changes in the population size, demographic characteristics of the population, technology and consumer preferences.
- Cyclical variations (C) are medium-term changes caused by circumstances which repeat in cycles and cause upward and downward swings, not of equal length, throughout the series. In business, cyclical variations are often correlated with the general business cycle of prosperity, recession, depression and recovery.

- Random or irregular variations (I) occur over short intervals and are unpredictable with no pattern to their behavior. These are disturbances due to 'everyday' unpredictable influences, such as weather conditions, illness, political unrest, theft, war and transport breakdowns.
- Seasonal variations (S) are short-term fluctuations that tend to repeat themselves within a year such as over days, weeks, months or quarters. Examples of seasonal variation are as follows:
 - Sales of ice cream will be higher in summer than in winter.
 - A doctor can expect a substantial increase in the number of flu cases every winter.
 - Shops might expect higher sales shortly before Christmas.
 - The telephone network may be heavily used at certain times of the day (such as mid-morning and mid-afternoon) and much less used at other times (such as in the middle of the night).

The classical model used by economists, also known as the multiplicative model, provides the clearest explanation of the four components that make up the time series and their relation to each other. This model assumes that each observation (y) is made up of a combination of the four components and is represented by the formula:

$$y = T \times C \times S \times I$$

This equation states that factors associated with each of these components can be multiplied together to provide the value of the observed dependent variable y .

Note: The trend component (T) is stated in the same units as y , while the remaining three components are expressed as percent adjustments. A value above 100 indicates an above average effect for the component, and a value below 100 indicates a below average effect.

For a series composed only of annual data, there is no seasonal component. In that case the time series model becomes:

$$y = T \times C \times I$$

Periodically reported data, such as monthly, quarterly, weekly, daily, etc., include the influence of all four components of the time series.

The process of division can remove or isolate any component of this model and is called **decomposition**.

Note: Analysis of cyclical and irregular influences on data is useful for describing past variations but because of their unpredictability, their value in forecasting is very limited. Instead, a number of business indicators are used to forecast cyclical turning points. Predicting cyclical and irregular variations requires techniques beyond the scope of this unit.

If we ignore the C and I components, since by definition they can't be predicted, the forecasting model will become:

$$\hat{y} = T \times S$$

Activity 7.1

The sales (R'000) for Turtle Toys have been analysed and the four components have been determined for the preceding four quarters. Find the missing values in the table below assuming a multiplicative model.

	Sales (y)	Trend (T)	Seasonal (S)	Cyclical (C)	Irregular (I)
Winter		1 000	50	107	101
Spring	820	1 100	70	105	
Summer	988			105	98
Autumn	2 623	1 300	200		97

7.2 Historigram

The standard graph to portray the behaviour of data over time is a line graph, known as a historigram.

1. Time is the independent variable (x) and is measured along the horizontal axis.
2. The variable of interest is the dependent variable (y) and is measured on the vertical axis.
3. Plot the time series values (on the vertical y-axis) against time (on the horizontal x-axis) as single points, and join the points by straight-line segments.

7.3 Time-series decomposition

7.3.1 Trend analysis (T)

The trend can be thought of as the core component of the time series model about which the other components, cyclical (C), seasonal (S) and irregular (I) variations fluctuate.

The objective is to enable the underlying movement of the data to be highlighted by making use of a straight line passing through the data with a positive or negative slope. Thus, a business sales trend will normally show whether sales are moving up or down (or remaining static) in the long term.

This component is found by identifying separate trend (T) values, each corresponding to a time point. Methods that can be used to extract a trend from a set of time series values are:

- method of least squares
- method of semi-averages
- method of moving averages.

Method of least squares

Steps

1. The observed time-series data are the dependent y -variable since this is the variable that we will want to predict.
2. The time variable is the independent x -variable. The time periods are translated into x -values by using a coding process. 1 represents the first time period, 2 the second time period, and so on until the final time period.
3. Having established the values for the x - and y -variables, the following formulae can be used to identify the trend line through the data:

$$\hat{y} = a + bx$$

Where \hat{y} is the trend value for a given time period

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

Where b = slope of the trend line

$$a = \frac{\sum y}{n} - b \frac{\sum x}{n}$$

Where a = y -intercept

4. To calculate the trend values (\hat{y}) with the trend line equation, substitute the appropriate x -code into the equation and compute the value of \hat{y} .
5. Plot the trend values \hat{y} together with the corresponding time periods on the time-series graph and draw a line through the points. This is the trend line. By extending this line, future values can be read from it.
6. To forecast future values using the equation, substitute the x with an appropriate code for the year of forecast (extrapolation) and calculate \hat{y} .

Example 7.1

The following table shows the income (R'000) of Super 10 Taxis by year.

Year	x -code	Income y	xy	x^2	\hat{y}
2001	1	28	28	1	29.0
2002	2	31	62	4	30.3
2003	3	34	102	9	31.6
2004	4	30	120	16	32.9
2005	5	35	175	25	34.2
	15	158	487	55	158

Hidden page

Hidden page

Hidden page

Activity 7.3

The following table shows the number of traffic tickets issued by the Alberton Traffic Department for the first six months of the year. Forecast the number of traffic tickets for July by making use of the method of semi-averages and portray the trend line on a graph.

Month	No. of tickets
January	120
February	120
March	100
April	90
May	130
June	150

Method of moving averages

This method attempts to remove variations from data by a process of averaging a fixed number of periods. The objective is to bring out the trend by eliminating any obscuring seasonal, cyclical or irregular fluctuations. One of the drawbacks is that values for some periods are lost at the beginning and end of the series.

A moving average is an artificially constructed time series in which the value for a given period is replaced by the mean of that value and the values of some number of preceding and succeeding time periods.

Steps

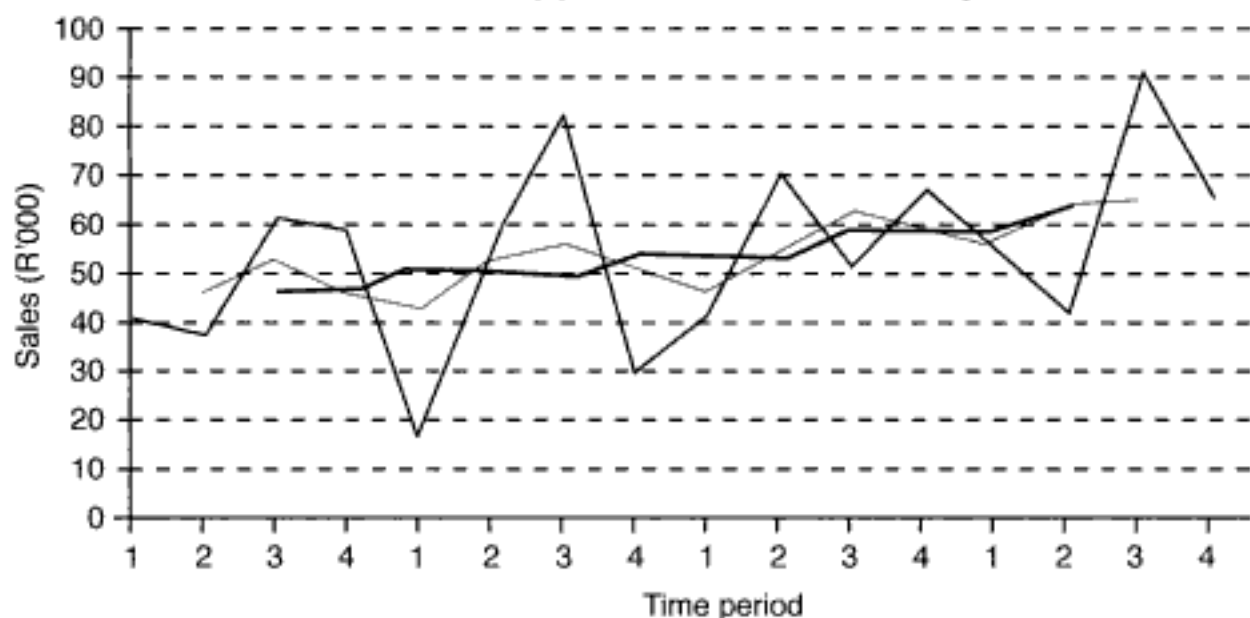
1. If you calculate an odd-numbered moving average (i.e. 3, 5 or 7), there will be a middle time point opposite which to record the answers. For example, in calculating a three-year moving average, you will start by adding the y -values for the first three years and dividing the answer by three. This answer will correspond with the middle of the second year. Move down one year and calculate the average for years two to four. This answer will correspond with the middle of year three. Complete the process for all the years.
2. If you calculate an even-numbered moving average (i.e. 2, 4, 6, or 8), the resulting averages will correspond between two time points. However, a trend value is required to coincide with a particular point in time; therefore an extra step is required to centre the average. This is done by calculating a moving average of two on the first moving average column.
3. Plot the moving averages on a graph with the original data to show how the time series is smoothed.

Hidden page

to calculate the average. Move down one quarter from the top and calculate the second value in the column: $(37 + 61 + 58 + 16) \div 4 = 43$. This answer corresponds with the position between the third and the fourth quarters.

These values do not correspond with the middle of a specific time period; therefore a centred column is required. The first value in this column is obtained by: $(49 + 43) \div 2 = 46$. This answer corresponds with the middle of the third quarter. Move down one quarter and calculate the second value: $(43 + 47.5) \div 2 = 45.2$. This answer corresponds with the middle of the fourth quarter.

Quarterly petrol sales for Jack's Garage



Activity 7.4

Construct a four-year and a five-year moving average to smooth the following time series and graph the results.

Year	y
2000	14
2001	20
2002	40
2003	30
2004	28
2005	42
2006	51
2007	25
2008	32

Hidden page

3. Calculate the trend value for each time unit within the forecasting period.
4. Multiply the trend value for each time unit with the seasonal index of that time unit.

Deseasonalising data

The influence of seasonality can be removed from a time series by dividing each original value in the series by the appropriate typical seasonal index for that period and then multiplying the result by 100. The result is known as deseasonalised data. Deseasonalised data are used if we wish to compare data across seasons to determine if an increase or decrease, irrespective of seasonal trends, has taken place.

Example 7.4

The quarterly income (R'000) of a soft drink company has been recorded for four years. The time period in date order is shown in column 1 and the actual sales (y) in column 2.

1. Quarterly data are given, therefore a four-quarterly moving average is determined and listed in column 3. The first value is: $(52 + 67 + 85 + 54) \div 4 = 64.5$. This answer corresponds with a position between the second and third quarters of 2004. The second value in the column is calculated by moving down one quarter: $(67 + 85 + 54 + 57) \div 4 = 65.8$. This answer corresponds with a position between the third and fourth quarters of 2004. By moving down one quarter at a time calculate the rest of the moving averages.
2. The moving average period (four-quarterly) is an even number therefore the values in column 3 must be centred. The first centred value is $(64.5 + 65.8) \div 2 = 65.2$. This answer corresponds with a position in the middle of the third quarter of 2004. The second value is calculated by moving down one quarter: $(65.8 + 67.8) \div 2 = 66.8$. This answer corresponds with a position in the middle of the fourth quarter of 2004. By moving down one quarter at a time, calculate the rest of the centred averages and enter them in column 4.
3. Obtain the percentage in column 5 by dividing the value in column 2 (actual income) by the corresponding value in column 4 (centred averages) and multiplying the result by 100. The first value is: $(85 \div 65.2) \times 100 = 130.4$. This corresponds with the third quarter of 2004.

Hidden page

Summary table:

Year	1	2	3	4	
2004			130.4	80.9	
2005	83.3	107.3	126.4	85.0	
2006	82.8	105.2	126.6	83.0	
2007	85.4	107.3			
Unadjusted	83.3	107.3	126.6	83.0	400.2
	× 0.9995	× 0.9995	× 0.9995	× 0.9995	
Typical index	83.3	107.2	126.5	83.0	400.0

8. Interpret index numbers: the influence of the season caused the sales during the quarter to be 16.7% lower than expected. The second quarter sales are 7.2% higher than expected, the third quarter sales are higher with 26.5% and the fourth quarter sales are lower than expected with 17% due to the influence of the season.

Forecasting:

1	2		6			
	y	Seasonal index	Deseasonalised y	x -code	xy	x^2
2004 1	52	83.3	62.4	1	52	1
2	67	107.2	62.5	2	134	4
3	85	126.5	67.2	3	255	9
4	54	83.0	65.1	4	216	16
2005 1	57	83.3	68.4	5	285	25
2	75	107.2	70.0	6	450	36
3	90	126.5	71.1	7	630	49
4	61	83.0	73.5	8	488	64
2006 1	60	83.3	72.0	9	540	81
2	77	107.2	71.8	10	770	100
3	94	126.5	74.3	11	1 034	121
4	63	83.0	75.9	12	756	144
2007 1	66	83.3	79.2	13	858	169
2	84	107.2	78.4	14	1 176	196
3	98	126.5	77.5	15	1 470	225
4	67	83.0	80.7	16	1 072	256
Total	1 150		1 150	136	10 186	1 496

- The deseasonalised data in column 6 are obtained by dividing the original data in column 2 by the appropriate typical index for that period. The first value in column 6 is: $(52 \div 83.3) \times 100 = 62.4$. This means that the income for the first quarter of 2004 would be R62.4 million if there had been no seasonal variation. The second value is: $(67 \div 107.2) \times 100 = 62.5$.
- To do a seasonalised forecast per quarter for 2008, code the original time period and use the method of least squares together with the sales data to obtain the quarterly trend equation: $\hat{y} = 61.59 + 1.21x$

Note: You can also use the deseasonalised data to calculate your trend values.

- Code the period you want to forecast and substitute the x in the trend equation with the appropriate code for that time period to obtain the trend value. Multiply each trend value with the seasonal index for that period.

$$\text{First quarter: } \hat{y} = 61.59 + 1.21(17) = 82.16 \quad 82.16 \times \frac{83.3}{100} = 68.44$$

Year	x -code	Trend forecast	Seasonal index	Seasonalised forecast
2008 1	17	82.16	83.3	68.43
2	18	83.37	107.2	89.37
3	19	84.58	126.5	106.99
4	20	85.79	83.0	71.21

Activity 7.5

The owner of a pizzeria recorded the number of pizzas sold during the past three weeks in order to determine the influence of the day of the week on the sales. Do a seasonalised forecast per day for week 5 and graph the original time series, the deseasonalised time series and the trend line.

Week	Monday	Tuesday	Wednesday	Thursday	Friday
1	12	18	16	25	31
2	11	17	19	24	27
3	14	16	16	28	25
4	17	21	20	24	32

TEST YOURSELF 7

1. Complete the following table assuming the classical multiplicative model:

	Trend (T)	C	S	Forecast
Winter	130	80	120	
Spring	132	90	100	
Summer	134	100	70	
Autumn	136	110	110	

2. Using the classical model, find the missing values:

	Trend	C	S	I	Sales
Winter	100	100		90	99
Spring	200		80	100	168
Summer	300	110		110	
Autumn	400	120	120		604.8
Total					

3. The total units of new government housing under construction for the past six years in the Gauteng Province are given below:

Year	Total units
2003	1 488
2004	1 014
2005	1 354
2006	1 474
2007	1 617
2008	1 666

Forecast the number of units that will be built during 2009 and 2010.

4. The data given below (in hundreds) were prepared by a marketing research agency for Radio UJ:

Year	Audience size
2000	31
2001	32
2002	33
2003	30
2004	29
2005	30
2006	28
2007	26

- a) Estimate the audience for the year 2008 using the method of least squares and graph the series.
 b) Estimate the audience for the year 2008 using the method of semi-averages.
5. a) Use the data given in the following table to smooth the time series by making use of a:
- two-year moving average
 - three-year moving average
 - four-year moving average.
- b) Graph the original data and the moving averages.
 c) Forecast the value for the year 2009.

2000	2001	2002	2003	2004	2005	2006	2007	2008
642	819	845	755	767	720	749	794	686

6. The number of operational nuclear power reactors in the world for the given years are listed in the following table:

Year	2002	2003	2004	2005	2006	2007
Number of reactors	83	99	108	110	105	104

Forecast how many nuclear plants will be operational during 2008.

7. The expenses (R'000) of Mr Hyde's Chemistry Research Lab is listed for the previous seven years:

2001	2002	2003	2004	2005	2006	2007
19 200	13 800	11 270	13 800	9 400	8 900	9 000

Forecast the expenses for 2008.

8. The table data represent the ultraviolet index for Durban during a holiday season:

1 Dec	2 Dec	3 Dec	4 Dec	5 Dec	6 Dec	7 Dec	8 Dec	9 Dec	10 Dec
9	4	10	10	9	8	8	10	10	9

Forecast the index for the next two days.

9. Metabolic rate is the rate at which the body consumes energy and is important in studies of weight gain, dieting and exercise. The metabolic rates, in calories per 24 hours, of a man who took part in a study of dieting, are recorded for seven weeks on the dieting program:

Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7
1 867	1 792	1 666	1 614	1 460	1 439	1 362

Forecast the metabolic rate for week 8.

10. The quarterly sales data (R'000) for Ajax washing power are provided below:

	Summer	Autumn	Winter	Spring
2001	10.4	11.8	8.5	7.5
2002	12.2	13.6	9.5	8.8
2003	13.5	13.1	10.4	9.7
2004	11.7	12.9	9.5	8.4
2005	13.7	15.0	10.9	10.1

- Do a seasonalised quarterly forecast for 2006.
- Deseasonalize the time series.
- Graph the time series, trend values and deseasonalised sales.

11. The sale of municipality houses to existing tenants is tabulated below:

Year	No. of houses sold		
	Jan–April	May–Aug	Sept–Dec
2005	43	80	60
2006	60	100	70
2007	80	120	90
2008	90	140	100

- Deseasonalize the time series.
- Forecast seasonalised house sales per term for 2009.

12. The electricity consumption (kilowatt per hour) for an engineering workshop is given in the following table:

	Jan–Feb	Mar–Apr	May–June	July–Aug	Sept–Oct	Nov–Dec
2006	245	220	190	185	200	225
2007	248	215	186	187	201	230
2008	250	225	189	188	198	226

- a) Deseasonalise the time series.
 b) Forecast seasonalised consumption per period for 2009.
13. The daily income (R'00) of a dry cleaner is tabulated below:

	Mon	Tues	Wed	Thurs	Fri	Sat
Week 1	27	23	29	28	37	55
Week 2	30	25	35	33	36	57
Week 3	32	28	34	34	40	54

- a) Deseasonalise the time series.
 b) Forecast seasonalised income for week 4.
14. The total expenditure (R'000) on part-time teaching by the statistics department is tabulated below:

	2005	2006	2007	2008
Jan–June	155	150	140	130
Jul–Aug	100	95	90	80

- a) Forecast the seasonalised expenditure per semi-annual for 2009.
 b) Deseasonalise the time series.

UNIT 8

Probability: basic concepts

The theory of probability grew out of the study of various games of chance using coins, dice, cards, lottery and gambling machines. Since then probability theory has been developed to determine uncertainties in our every day lives as well.

After completion of this unit you will be able to:

- define probability
- describe the classical, empirical and subjective approaches to probability
- understand the meaning of basic terms used in the probability theorem
- apply the properties of probability
- calculate probabilities using the rules of addition and multiplication
- use the counting rules.

In each of the following questions is an implied condition of uncertainty:

- Is there a link between second-hand smoking and asthma in young children?
- What are my chances of getting that new job?
- What is the estimated influence of the Aids epidemic on population growth?
- What is the chance that it will rain today?

We make decisions in the face of uncertainty. That means, facing situations where it is possible that things could turn out in different ways, but we simply do not know how probable each event or outcome is. Our need to cope with this risk or the 'chance that it will happen' leads us to the study and use of probability theory.

Inferential statistics involves using statistics obtained from a sample to make estimates and decisions concerning the entire population. We can never be certain that our decisions are correct, but to assess how good they will be, we need to know how to measure 'chance' and 'probability'. The science of measuring 'uncertainty' is called probability.

Note: Probability describes the relative possibility (chance or likelihood) that an event will occur.

8.1 Language of probability

An *experiment* or *investigation* is an action that generates the uncertain outcomes to which we will assign probabilities.

A particular result of an experiment is an *outcome*.

A *sample space* of a random experiment is a list of all the possible outcomes of the random experiment.

The individual outcomes in a sample space are called *simple events*. An *event* is a collection of one or more outcomes of an experiment.

Example 8.1

A family has three children. Their blood type can either be O or A. The sample space for the blood types of the three children contains eight possible outcomes:

[OOO, AOO, OAO, OOA, AAO, AOA, OAA, AAA]

Each of these outcomes determines a simple event.

1. The event that exactly one of the children is of blood type A
= [AOO, OAO, OOA]
2. The event that at most one child has blood type A
= [OOO, AOO, OAO, OOA]
3. The event that all children have the same blood type = [AAA, OOO]

Activity 8.1

The type of transmission – automatic (A) or manual (M) – is recorded for each of the next two cars purchased from a certain dealer.

1. What is the random experiment?
 2. What is the sample space?
 3. List the outcomes in each of the following events:
 - a) that at least one car has an automatic transmission
 - b) that exactly one car has an automatic transmission
 - c) that neither car has an automatic transmission.
-

8.2 Approaches to assigning probabilities

Each outcome in a sample space has a probability. Each event has a probability. The method you will use to calculate a probability depends on the approach you use.

8.2.1 Classical approach

Classical probability is used in situations where the outcomes of the experiment are equally likely. The probability of an outcome or event happening is computed by dividing the number of favorable outcomes (or successes) by the number of possible outcomes.

$$P(E) = \frac{\text{number of successes}}{\text{total number of outcomes}}$$

$P(E)$ is read as the probability that event E will occur.

Steps

1. Identify the event (or success).
2. Find the number of successes.
3. Find the total number of outcomes in the experiment.
4. Divide the number of successes by the total of possible outcomes.
5. Interpret the probability.

Example 8.2

In rolling a fair die once, each of the possible outcomes in the sample space (1, 2, 3, 4, 5, 6), has an equal chance of occurring. In calculating the probability of the event obtaining an even number in one roll of the die, your number of successes is (2 or 4 or 6).

$$P(\text{even number}) = \frac{3}{6} = 0.5$$

Activity 8.2

In drawing a card from a deck of 52 cards, what is the probability that it will be an ace?

8.2.2 Empirical probability

This approach is based on relative frequencies. The probability of an event happening is determined by observing what proportion of the time similar events happened in the past or if the experiment is repeated many times under identical conditions.

$$P(E) = \frac{\text{number of times the event occurred}}{\text{total number of observations}}$$

As you increase the number of times an experiment is repeated, the empirical probability of an event approaches the classical probability of the event.

Note: Chance behaviour is unpredictable over the short run but has a regular and predictable pattern in the long run.

Steps

1. Identify the event (or success).
2. Find the frequency of the event. That is the number of times the event occurred in the experiment or in the past.
3. Find the total number of outcomes in the experiment.

Hidden page

Hidden page

Example 8.5

- a) If you flip a coin, the possible outcome is heads or tails.
 The event of obtaining heads is $P(H) = \frac{1}{2}$
 The event of obtaining tails is $P(T) = \frac{1}{2}$
 If you add the probabilities of the two possible outcomes, the total = 1
- b) A restaurant wants to determine the probability that its manager is going to reject the next delivery of fresh oysters from a supplier. Records show that the supplier sent the restaurant 90 batches of oysters in the past, and the manager rejected 10 of them.
 $P(\text{rejecting next batch}) = \frac{10}{90} = 0.11$
 $P(\text{accepting next batch}) = \frac{80}{90} = 0.89$

If you add the probabilities of the two possible outcomes, the total is 1.

3. **The complement rule:** The complement of an event E is an event \bar{E} that does not occur. That means all the outcomes in the sample space that do not belong to event E . The sum of the probabilities assigned to all the possible outcomes in a sample space is equal to 1. If the probability of occurrence of event E is $P(E)$ and the probability that event E will not occur is $P(\bar{E})$, then

$$P(E) + P(\bar{E}) = 1 \text{ and } P(\bar{E}) = 1 - P(E)$$

Despite its simplicity, the complement rule can be very useful. The task of finding the probability that an event of interest will *not* occur is sometimes easier or less time-consuming than finding the probability that it will occur.

Example 8.6

If the probability of completing a job is 0.8, the probability of not completing the job is:

$$P(\text{not completing the job}) = (1 - 0.8) = 0.2$$

Activity 8.5

The probability that a typist will make at most five mistakes is 0.64. What is the probability that she will make more than five mistakes?

Activity 8.6

We all know that fruit is good for us and that we don't eat enough. In a recent study done among a random sample of 75 teenage boys, the following information was collected:

Fruit servings per day	Number of boys	% of boys
0	20	27
1	15	20
2	15	20
3	12	16
4	8	11
5	5	6
	75	100

1. What is the probability of teenage boys eating three fruit servings per day?
2. What is the probability of teenage boys not eating three fruits per day?
3. What is the probability that teenage boys will eat at least one fruit per day?
(Note: 'At least one' means one or more)

8.4 Forming new events

Once some events have been specified, there are several useful ways of manipulating them to create new events known as compound events.

8.4.1 Union of two events (A or B)

The union of events A and B is represented by the symbol $A \cup B$ and merges all the outcomes in event A with those in event B. An outcome is only listed once in the combined event if it appears in both A and B.

8.4.2 Intersection of two events (A and B)

The intersection of two events is represented by the symbol $A \cap B$ and contains all the outcomes that events A and B share in common – this is known as a joint probability.

Example 8.7

In an experiment consisting of selecting one card from a deck of playing cards, you may be interested in some simple possible events, for example the card selected is:

- A = the queen of hearts
- B = a red card
- C = a king
- D = a face card
- E = an ace

Some compound events that you can form from the simple events listed above:

- B or E = you are interested in drawing either one of the 26 red cards or any one of the possible four aces of which two are red – that is a total of 28 possible outcomes that will satisfy your event. Remember you can list an outcome only once.
- A or C = you are interested in the possibility that the card you select will either be the queen of hearts or a king – that can be any one of five possible outcomes.
- B and C = you are interested in the possible outcomes that will result in a red, king card – it can be one of two possible outcomes. Only two of the king cards are red.

8.4.3 Display events graphically

It is sometimes useful to draw a picture of events in order to visualise relationships and understanding probability.

Diagrams used to display events are known as Venn diagrams.

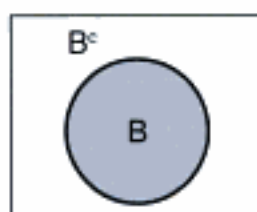
The collection of all possible outcomes, that is the sample space, is shown as the interior of a rectangle with the various events drawn as circles inside the rectangle. Each event is a subset of the sample space.

Example 8.6 is used to illustrate.

Only one event is displayed

Venn diagram for event B : drawing a red card

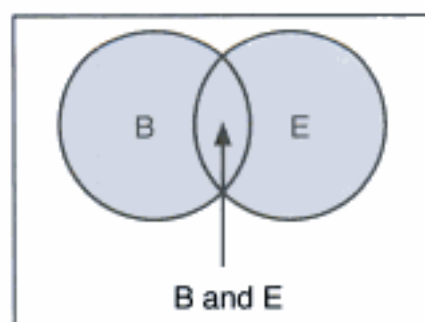
B^c : not drawing a red card – compliment rule.



The union of two events – B or E : $B \cup E$

The total area in the Venn diagram displays the union of two events. It consists of all the outcomes in either event B or event E or both. The area where B and E overlaps is where both B and E occur.

B or E = you are interested in drawing either a red card or any one of the four aces. There are 26 red cards (B) and four aces (E), of which two are also red (B and E). There are 28 possible outcomes that will satisfy your event.

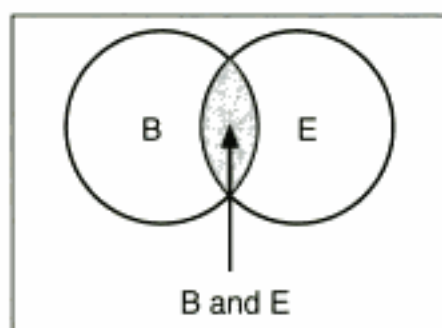


Intersection of two events – B and E : $B \cap E$

The shaded area in this Venn diagram displays the intersection of two events B and E.

This compound event is defined by the condition that both event B and event E occur and consists of all outcomes common to both event B and event E.

B and E = you are interested in drawing a red ace. There are 26 red cards (B) and four aces (E), of which two are also red (B and E). There are two possible outcomes that will satisfy your event.



Union of two mutually exclusive events: A or C

Two events are said to be mutually exclusive if when one occurs, the other cannot occur at the same time. The two events are disjoint and have no outcomes in common.

A or C = you are interested in the possibility that the card you select will either be the queen of hearts or a king – that can be any one of five possible outcomes.



8.5 Probability rules for compound events

We can use various rules of probability to compute the probabilities of the more complex, related events.

1. If an event (E) is made up by two (or more) simple events (A and B), the probability $P(E)$ can be formed either as:
 - the union of two or more events – all the outcomes that make up the two events
 - the intersection of two events – the outcomes that fulfill the conditions for both events.
2. Two or more events are **mutually exclusive** if the occurrence of one event means that none of the other events can occur at the same time. An outcome can belong to event A **or** to event B but not to both.

If you flip a coin, you can either have heads or tails. Both can't happen at the same time!

3. Two events are **independent** if the occurrence of one is in no way affected by the occurrence of the other; that is, they are unrelated.

If you flip two coins and you obtained heads on the one, it will have no influence on the outcome of the second flip.

4. If there is a particular relationship between events such that the occurrence of one event affects the occurrence of the second event, the events are **dependent** and the probability attached to the occurrence of such events is known as **conditional probability**.

8.5.1 Addition rules

Events are combined if we want to find the probability of one event **or** another occurring.

Special rule of addition

To apply the special rule of addition, the events must be **mutually exclusive** - the two events are disjoint and therefore cannot occur simultaneously.

This rule states that the probability of event A **or** event B occurring in a given **single outcome** equals the sum of their probabilities. This is known as the union of P(A) with P(B).

$$P(A \text{ or } B) = P(A) + P(B)$$

Example 8.8

We all know that fruit is good for us and that we don't eat enough. In a recent study done among a random sample of 75 people, the following information was collected:

Fruit servings per day	Number of people
0	20
1	15
2	15
3	12
4	8
5	5
Total	75

- The probability that a selected person eats two fruits per day is $\frac{15}{75} = 0.20$.
The probability that a selected person will eat three fruits per day is $\frac{12}{75} = 0.16$.
The two events are mutually exclusive because if the person eats exactly two fruits per day he cannot eat exactly three fruits as well.
- The probability that a selected person will eat two or three fruits per day is $P(2 \text{ or } 3) = P(2) + P(3) = 0.20 + 0.16 = 0.36$.

- The probability that a randomly selected person will eat at least four fruits per day is ...

Activity 8.7

If the probabilities are 0.05, 0.14, 0.17, 0.33, 0.20 and 0.11 that a wine tasting will rate a new Shiraz as very poor, poor, fair, good, very good or excellent, what are the probabilities that the Shiraz will be rated as:

1. very poor or poor
2. good, very good or excellent.

General addition rule

If the events are *not* mutually exclusive, it means that event A may occur or event B may occur, or both A and B may occur in a **single outcome**.

This rule states that the probability that either event A **or** event B occurs equals the probability that event A occurs plus the probability that event B occurs minus the probability that both occur.

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

To avoid double counting the probability of the outcomes that fulfill the conditions for both events, $P(A \text{ and } B)$ is subtracted from the sum of the probability of A and B.

Note: In this rule, $P(A \text{ and } B)$ denotes the probability that A and B both occur in the same observation. In the multiplication rule $P(A \text{ and } B)$ denotes the probability that event A occurs on one trial followed by event B on another trial.

Example 8.9

The probability that a person stopping at a petrol garage will ask to have his tires checked is 0.12, the probability that he will ask to have his oil checked is 0.29 and the probability that he will be asked to have both checked is 0.07. What is the probability that a person stopping at this garage will ask to have:

- either his tires or his oil checked?

$$P(T \text{ or } O) = P(T) + P(O) - P(T \text{ and } O)$$

$$= 0.12 + 0.29 - 0.07 = 0.34$$
- neither his tires nor his oil checked?

$$1 - P(T \text{ or } O) = 1 - 0.34 = 0.66$$

Activity 8.8

There are two secretaries in the office. The probability that the one secretary will be absent on any given day is 0.08 and the probability that the other one will be absent on any given day is 0.07. The probability that both will be absent is 0.02. What is the probability that on a given day:

Hidden page

Conditional probability is the probability of an event B occurring **given** that event A has already occurred. The probability of the second event B is affected by the outcome of the first event A.

$$P(A \text{ and } B) = P(A) \cdot P(B|A)$$

Conditional probability is denoted by $P(B|A)$ and is read as the 'probability of B, given A happened'.

Example 8.11

You are not aware of it, but in a case of wine bought, five of the twelve bottles are bad.

The given table lists all the possible outcomes in the experiment together with the probability for each possible outcome.

G = good B = bad G_1G_2 = first bottle good and second bottle good

1. To calculate $P(G_1G_2)$:

There are seven good bottles in the case of 12 $\therefore P(G_1) = \frac{7}{12}$.

The probability of G_2 if the first bottle is good (G_1): $P(G_2|G_1) = \frac{6}{11}$ (Remember that if the first bottle selected is good, there are only 11 bottles left in the box and only six good ones. The condition is: if the first bottle is good.)

All the possible outcomes	Probability for each outcome
G_1G_2	$\frac{7}{12} \times \frac{6}{11} = \frac{42}{132}$
G_1B_2	$\frac{7}{12} \times \frac{5}{11} = \frac{35}{132}$
B_1G_2	$\frac{5}{12} \times \frac{7}{11} = \frac{35}{132}$
B_1B_2	$\frac{5}{12} \times \frac{4}{11} = \frac{20}{132}$
Total	$\frac{132}{132} = 1$

If you were to select two bottles from the case, what is the probability that:

2. both bottles are bad?

$$P(B_1B_2) = \frac{5}{12} \times \frac{4}{11} = \frac{20}{132}$$

3. one of the two bottles is bad?

There are two possibilities that will give you this answer.

Note: The probability of an event is the sum of all the possibilities that will give you the answer.

$$P(G_1B_2) + P(B_1G_2) = \left(\frac{7}{12} \times \frac{5}{11}\right) + \left(\frac{5}{12} \times \frac{7}{11}\right) = \frac{70}{132}$$

Example 8.12

A certain testing apparatus has two batteries. The probability that the first battery will run down is 0.3 and the probability that both batteries will run down is 0.06. If the first battery is found to be flat, what is the probability that the second battery will be flat?

$$P(F_1 \text{ and } F_2) = P(F_1) \times P(F_2 \setminus F_1)$$

$$0.06 = 0.30 \times P(F_2 \setminus F_1)$$

$$\therefore P(F_2 \setminus F_1) = \frac{0.06}{0.30} = 0.20$$

This means that 20% of the time, if the first battery is flat, the second battery will also be flat.

Activity 8.10

A medical researcher has discovered a new test for tuberculosis. Experimentation has shown that the probability of a positive test is 0.82, given that a person has tuberculosis. The probability is 0.04 that the test registers positive and that the person does not have tuberculosis. Assuming that in the general population the probability that a person has tuberculosis is 0.20, what is the probability that a person chosen at random will:

1. have tuberculosis and a positive test?
2. not have tuberculosis?
3. have a positive test but not have tuberculosis?

8.5.3 Calculating probabilities using a contingency table

A two-way contingency table or cross-tabulation table lists the frequency of each combination of the values of two variables.

Example 8.13

Equity among the judges

The female judges in a certain province recently lodged a complaint about the most recent round of promotions. An analysis of the relationship between gender and promotion were undertaken with the joint probabilities given in the following table:

	Promoted	Not promoted	Total
Female	0.03	0.12	0.15
Male	0.17	0.68	0.85
Total	0.20	0.80	1.00

$$P(\text{female}) = 0.15$$

$$P(\text{promoted}) = 0.20$$

$$P(\text{female and promoted}) = 0.03$$

$$P(\text{male}) = 0.85$$

$$P(\text{not promoted}) = 0.80$$

$$P(\text{male and promoted}) = 0.17$$

$$P(\text{female and not promoted}) = 0.12$$

$$P(\text{male and not promoted}) = 0.68$$

1. What is the rate of promotion among female judges?

From this table you can see that of all the judges in that province, only 0.15 or 15% are female. The rest, 85%, are male. To calculate the probability of being promoted if you are a female, the following formula will be appropriate:

$$P(F \text{ and } P) = P(F) \cdot P(P|F)$$

$$0.03 = 0.15 \cdot P(P|F)$$

$$P(P|F) = 0.20$$

2. What is the rate of promotion among male judges?

$$P(M \text{ and } P) = P(M) \cdot P(P|M)$$

$$0.17 = 0.85 \cdot P(P|M)$$

$$P(P|M) = 0.20$$

3. Is an accusation of gender bias reasonable?

No, because the same proportion of the females are promoted as the males.

Activity 8.11

The table shows the results of a study on 102 children in which a child's IQ was examined and the presence of a specific gene was found in the child.

	Gene present	Gene not present	Total
High IQ	33	19	52
Normal IQ	39	11	50
Total	72	30	102

Determine the probability that a child has:

1. a high IQ and the given gene
2. a normal IQ without the gene
3. the gene
4. a normal IQ
5. a high IQ, given that the child has the gene
6. the gene, if his IQ is normal.

8.5.4 Tree diagrams

Another useful method of calculating probabilities if there are several stages or trials in the experiment is to use a probability tree. All the possible outcomes of the experiment are represented by the branches of the tree.

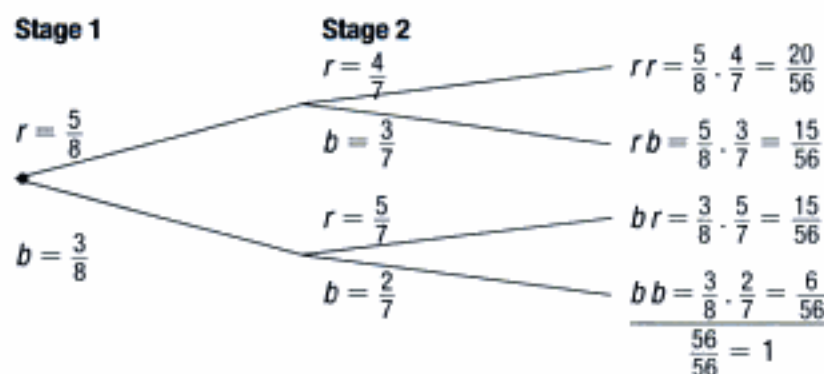
Steps

1. Plot a dot on the left to represent the root of the tree.
2. Construct a column for each trial.
3. Start on the left and determine the possibilities for the first trial which forms the branches of the tree in the first column.
4. Branches grow from each of the original branches, representing the possibilities for the second trial. The *second stage* is based on the choice made in the first stage. Determine if the outcomes are dependent or independent.
5. The branches of the tree are weighted by probabilities; therefore show the probabilities for each event on the branches.
6. List all the outcomes together with the joint probability for each combined outcome.
7. Add the probabilities. Because the tree represents the sample space of the experiment, the sum of the probabilities should equal 1.

Example 8.14

A bag contains five red balls and three black balls. Two balls are drawn from the bag. Construct a probability tree to list all the possible outcomes together with each outcome's probability.

1. The **first stage** of the tree consists of the possibilities in the first draw. There are only red and black balls in the bag, therefore if you draw one ball from the bag it can either be red or black. These possible outcomes are represented by the branches of a tree.
2. If the ball in the first draw was red, there are still four red balls and three black balls in the bag, therefore the second ball you draw can either be red or black. If the first ball was black, there are still five red and two black balls in the bag, therefore the second ball you draw can either be red or black.
3. The probabilities alongside each possibility must be calculated. In the first stage if you draw a ball, the chance that it is red is $\frac{5}{8}$. The chance that the first ball is black is $\frac{3}{8}$. Please note that the first ball can either be red or black. Only one of the two possibilities can happen.
4. If the first ball is red, there are only four red ones left and seven balls in the bag, therefore the chance that the second one is red is only $\frac{4}{7}$. But if the second one is black the probability is $\frac{3}{7}$.
5. If the first ball was black, the chance that the second one is red is $\frac{5}{7}$ or the second one can also be black with an associated probability of $\frac{2}{7}$.



- If you add the probabilities of all the possible events, the total must be one.
- From the tree diagram, the probability of drawing two red balls is $\frac{20}{56}$.

The probability of a red and a black ball is $\frac{15}{56} + \frac{15}{56} = \frac{30}{56}$.

There are two possible outcomes that result in a red and a black ball. A probability of an event is the sum of the probabilities of all the possibilities that can result in the required event.

Activity 8.12

Approximately 10% of people are left-handed. If two people are selected at random, use a probability tree to determine the probability that:

- both are right handed
 - both are left-handed
 - one is right-handed and the other left-handed.
-

8.6 Counting the possibilities

Probability is based on the number of successes and the possible number of outcomes that make up the numerator and denominator. A collection of rules for counting the number of outcomes that can occur for a particular experiment can be used.

8.6.1 Multiplication rule of counting

The multiplication counting rule (mn rule) states that if there are m ways in which event A can happen, and n ways in which event B can happen, then there are m times n ways in which both can happen. This rule can be extended if there are more events.

$$m \times n \times \dots$$

Example 8.15

A computer password is to be made up consisting of four alphabetical characters. How many different computer passwords can be designed if repetition of letters is allowed?

There are four slots to be filled: a slot for each number in the password.
Thus there are $26 \times 26 \times 26 \times 26 = 456\,976$ computer passwords.

Activity 8.13

If a restaurant menu had a choice of three salads, six main dishes and six desserts, how many different possible dinners can be ordered?

8.6.2 Permutation rule

The permutation rule is used to determine the number of ways to arrange n distinct objects taking them x at a time in a specific order.

$${}^n P_x = \frac{n!}{(n-x)!}$$

Note: $n!$ (Pronounced as n factorial) is the product of the whole numbers from n downwards to 1. For example $4! = 4 \times 3 \times 2 \times 1$ and $0! = 1$. The factorial key is available on most calculators.

Example 8.16

Suppose we need to select a group of three people from a larger group of 10. They are to fill the roles of chairperson, secretary and treasurer in a committee. The number of possible ways of filling these roles is:

$${}_{10} P_3 = \frac{10!}{(10-3)!} = 720$$

Activity 8.14

Assume there are five carriages that need to be unloaded at a dock but there is only enough time left in the day to unload three of them. Since the goods in each of the carriages are needed by customers, the order of unloading is important. In how many ways can three of the five carriages be unloaded in first, second and third order?

8.6.3 Combination rule

The combination rule is used to determine the number of ways to select x objects from a larger set of n objects without regard to the order in which the objects are selected. For example ABC is considered the same selection as BCA or CBA. The number of combinations of n objects taken x at a time is:

$${}^n C_x = \frac{n!}{x!(n-x)!}$$

Example 8.17

A group of seven mountain climbers wishes to form a mountain climbing team of five. How many different teams could be formed?

$${}^7 C_5 = \frac{7!}{5!(7-5)!} = 21$$

Activity 8.15

You are given a list of 10 books and you are to read four of them. How many possible combinations of four books are available from the list of ten?

TEST YOURSELF 8

- There are six balls of the same size in a box. Two are red, three are blue and one is yellow. If you draw a ball from the box, what is the probability that:
 - a red ball will be selected?
 - the ball will not be yellow?
 - the ball will be red or yellow?

- This table shows the blood type of a randomly chosen person in South Africa:

Blood type	A	O	B	AB
Probability	0.40	0.45	0.11	?

- What is the probability that a randomly chosen person has type AB blood?
 - If you have type B blood and can receive blood transfusions from people with blood types O and B, what is the probability that a randomly chosen donor can donate blood to you?
- The table below shows the results of a survey in which 500 adults were asked why they don't always eat healthy foods:

Reason	Number responding
No time to cook	175
Not available as take-aways	95
High cost	85
Poor taste	60
Hard to find	55
Confusion about nutrition	30

- Find the probability of a randomly selected adult who doesn't always eat healthy foods because he or she has no time to cook or is confused about nutrition.
 - Find the probability of a randomly selected adult who feels that healthy foods have poor taste or are hard to find.
- What is the probability that an even number will result from one roll of a die?
 - If you draw a Smartie at random from a box of Smarties, you can draw one of six possible colours.

Colour	Blue	Green	Brown	Orange	Red	Yellow
Probability	?	0.16	0.13	0.20	0.13	0.14

- What is the probability of drawing a blue Smartie?
- What is the probability that you will not draw a brown Smartie?

- c) What is the probability that the Smartie you draw will either be yellow, orange or red?
6. The probability that a car owner in a certain income bracket will drive a Ford is 0.34 and the probability that he will drive a Toyota is 0.08. Find the probabilities that such a person will:
- not drive a Ford
 - drive a Ford or a Toyota
 - drive neither a Ford nor a Toyota.
7. The probability that a student at the university has hearing problems is 0.09; the probability that a student has eyesight problems is 0.15. The probability that a student will have a hearing and an eyesight problem is 0.01. What is the probability that a randomly selected student will have a hearing or an eyesight problem, but not both?
8. A welfare worker is studying the residents of a certain retirement community. She finds that 20% of the residents receive disability payments and 85% receive retirement incomes. 15% receive both disability and retirement incomes. If a resident is randomly chosen, what is the probability that the person receives a disability payment or retirement income?
9. In a certain lottery, the probability of drawing a number divisible by two is $\frac{1}{2}$, divisible by three is $\frac{1}{3}$ and divisible by six is $\frac{1}{6}$. What is the probability of drawing a number that is divisible by either two or three?
10. What is the probability of drawing either a heart or an ace from a deck of 52 playing cards?
11. An analysis of students' records at a university revealed that 45% of the students have an average C-symbol and 25% have jobs. 10% of the students have jobs and an average C-symbol. What is the probability that a student selected at random will have an average C-symbol or have a job?
12. Of 100 individuals who applied for a lab technician position with a large firm during the past year, 40 had some prior work experience and 30 had a professional certificate. However, 20 of the applicants had both work experience and a professional certificate. Determine the probability that a randomly selected candidate had:
- either work experience or a certificate
 - either work experience or a certificate but not both
 - a certificate, given that she had experience.
13. John is interviewed for a job at Karco. The probability that, after the interview, he will want the job is 0.88. The probability that Karco will want him is 0.45. The probability that he will want the job if Karco wants him is 0.92.
- What is the probability that John will want the job and that Karco will want him?
 - What is the probability Karco will want John if John wants the job?
14. The probability that a customer selects a pizza with mushrooms or pepperoni is 0.43, and the probability that the customer selects mushrooms only is 0.32. If the probability that he selects pepperoni only is 0.17, find the probability of the customer selecting both items.
15. In a sample of 1 000 people, 120 are left handed. Two unrelated people are selected at random from the sample. Find the probability that:

- a) both people are left-handed
 - b) neither person is left-handed
 - c) at least one of the two people is left-handed.
16. Out of 100 cars that start in the Grand Prix race, only 60 finish. The Total team in the race enters two cars. What is the probability that:
- a) both cars will finish?
 - b) neither of the two will finish?
17. If 18% of all South Africans are underweight, find the probability that if two citizens are selected at random, both will be underweight.
18. Approximately 9% of men have a type of colour blindness that prevents them from distinguishing between red and green. If three men are selected at random, find the probability that all three will have this type of red-green colour blindness.
19. The probability that a person has type O⁺ blood is 38%. Three unrelated people are selected in a random sample. Find the probability that:
- a) all three have type O⁺ blood
 - b) none of the three has type O⁺ blood
 - c) at least one has type O⁺ blood.
20. Students take two independent tests. 30% of them pass test A and 60% pass test B. Find the probability that a student selected at random passes:
- a) both tests
 - b) only test A
 - c) only one test.
21. Ten students are being interviewed for appointment to the students' council. Six of them are female and four are male. If two are selected at random for a newspaper interview, what is the probability that:
- a) at least one is female?
 - b) one female and one male are selected?
22. A person owns a collection of 30 CDs of which five are classical music. If two CDs are selected at random, find the probability that both are classical music.
23. A doctor gives a patient a 60% chance of surviving bypass surgery after a heart attack. If the patient survives the surgery, he has a 50% chance that the heart damage will heal. Find the probability that the patient survives the surgery and the heart damage will heal.
24. The probability that Jack parks in a disabled parking zone and gets a parking fine is 0.06. The probability that Jack cannot find a legal parking space and has to park in the disabled parking is 0.20. On Monday, Jack arrives at the shopping centre and has to park in the disabled parking zone. Find the probability that he will get a parking fine.
25. A batch of 10 calculators has three defective calculators. What is the probability that a sample of three calculators will have:
- a) no defective calculators?
 - b) all defective calculators?
 - c) at least one non-defective calculator?

Hidden page

What is the probability that a randomly selected athlete:

- is male?
 - is female?
 - tested positive?
 - is female and tested negative?
 - is either male or tested positive?
 - if he is tested positive, he is male?
 - if it is a female, she tested positive?
30. A boutique owner buys from three companies: A, B and C. Last month's purchases are shown in the table below:

Product	A	B	C
Dresses	24	18	12
Trousers	13	36	15

If an item is selected at random, what is the probability that it:

- was purchased from company A or is a dress?
 - was purchased from company B or company C?
 - is a trouser or was purchased from company A?
31. Consumers were surveyed on the number of visits to a new shopping centre and if the centre was conveniently situated. The recorded data are summarised in the following table:

Visits	Convenient	Not convenient
Often	60	20
Occasional	25	35
Never	5	60

What is the probability that a randomly selected consumer:

- visits the centre often and finds it convenient?
 - if it is convenient, visits it occasionally?
 - never visits it?
 - finds the centre not convenient?
32. Students are engaged in various sports in the following proportions:
- | | |
|-----------------------------|------------------------------|
| Rugby, 30% of all boys. | Cricket, 20% of all boys. |
| Soccer, 20% of all boys. | Both rugby and cricket, 5%. |
| Both rugby and soccer, 10%. | Both cricket and soccer, 5%. |
| All three sports, 2%. | |

If a student is randomly chosen, use a Venn diagram to calculate the chance that:

- he will play at least one sport
- he will be a rugby player or a soccer player
- he does not play any sport.

33. Common sources of caffeine are coffee, tea and cola drinks. Suppose that 55% of students drink coffee, 25% drink tea and 45% drink cola. Additional to that, 15% drink both coffee and tea, 5% drink all three, 25% drink both coffee and cola and 5% drink only tea. Draw a Venn diagram showing this information and determine:
 - a) what percentage of students drink only cola
 - b) what percentage drink none of these beverages.
34. There are four blood types, A, B, AB and O. Blood can also be Rh^+ or Rh^- . Finally, a blood donor can be classified as either male or female. In how many different ways can a donor's blood be labelled?
35. Four wires (red, green, blue and yellow) need to be attached to a circuit board by a robotic device. The wires can be attached in any order and the supervisor wishes to determine which order would be fastest for the robot to use. How many possible sequences of assembly must be measured?
36. A research biologist is studying the effects of fertiliser type, temperature at time of application, and water treatment after application on green beans. She has four fertiliser types, three temperature zones and three water treatments to test. Determine the different number of plots she needs in order to test each fertiliser type, temperature range and water configuration.
37. The access code for your office's security system consists of four digits. Each digit can be 0 through 9. How many access codes are possible if each digit can be used only once and not repeated? What will the answer be if each digit can be repeated?
38. A food processing plant packages all its food in clear plastic that is sealed. The quality control for the packaging process checks for three items: (1) the weight shown is correct, (2) the label is correct and (3) the package is properly sealed. These processes can be done in any order. In how many different ways can a package be cycled through the three inspection stations?
39. If a coin is tossed four times, how many different outcomes are possible?
40. Space shuttle astronauts each consume an average of 3 000 calories per day. One meal normally consists of a meat dish, a vegetable dish and a dessert. The astronauts can choose from 10 meat dishes, eight vegetable dishes and 13 desserts. How many meal combinations are possible?
41. A mail-order company sells eight different books. As part of a special promotion, customers may select three different books to make up a package. How many different packages are possible?
42. There are 15 qualified applicants for five trainee positions in a fast food management program. How many different groups of trainees can be selected?
43. A food technologist must select three tests to perform on ice cream. He has a choice of seven tests. In how many ways can he perform three different tests?
44. A new drug is in the test phase: the first phase involves five volunteers and the objective is to test the safety of the drug. If eight volunteers are available and five of them are to be selected, how many different combinations of five volunteers are possible?
45. The CEO of a research centre has to reduce the management staff from 10 to seven. He wants to get rid of the eldest three. How many possible arrangements are there of the management staff in order of age?

46. The Big Triple at the local race track consists of picking, in the correct order, the first three horses in the ninth race. How many possible Big Triple outcomes are there if the ninth race is run by 12 horses? What is the probability that your ticket will be a winning ticket?
 47. A rugby team must schedule a game with each of three other teams. There are five dates available for games. How many different schedules can be arranged? What is the probability that it will be scheduled on one specific day?
 48. Suppose that 60 of 200 students have statistics as a subject, 40 have accountancy as a subject and 25 have both subjects. Portray the given data using a Venn diagram and answer the following questions:
 - a) How many students have only statistics as a subject? Calculate the probability of having only statistics as a subject.
 - b) How many have only accountancy? What is the probability of taking only accountancy as a subject?
 - c) How many have statistics or accountancy or both? Calculate the probability attached to this answer.
 - d) How many students have neither of the two subjects? What is the probability of taking neither of the two subjects?
-

UNIT 9

Probability distributions

This unit introduces probability distributions and shows how they are used in practice.

After completion of this unit you will be able to:

- define a probability distribution
- distinguish between discrete and continuous random variables
- find the probability for a binomial investigation
- find the probability for a Poisson investigation
- find the probabilities for a normally distributed variable by transforming it into a standard random variable.

In statistical investigations involving chance, it is difficult to predict the exact value of a variable and so it is known as a **random variable**.

A **probability distribution** is a listing of all the possible outcomes a random variable can take.

The probability distribution of a random variable (x) assigns a probability to each of the possible values that the random variable can take. The sum of all the probabilities is 1.

Probability distributions are classified as either **discrete** or **continuous**, depending on the random variable.

- A random variable is **discrete** if it can assume a countable number of possible values like 0, 1, 2, 3, etc.
- A **continuous** random variable has an infinite number of possible values; that is, it can take on any value over a given interval of values.

It is important that you can distinguish between discrete and continuous random variables because different statistical techniques are used to analyse each.

9.1 Discrete probability distributions

Constructing a discrete probability distribution

Steps

1. List all the possible outcomes of an investigation in a frequency distribution.

Hidden page

Characteristics

- The investigation must consist of n identical trials.
- The trials are independent. That is, the outcome of one trial does not affect the outcome of any other trial.
- Each trial has one of two possible mutually exclusive outcomes: success or failure.
- Each trial has the same probability (π) of a 'success'.
- The probability of a failure is denoted by $(1 - \pi)$.
- The random variable (x) is the number of successes in the n trials of the investigation.
- The probability distribution of x is given by:

$$P(x) = \left(\frac{n!}{x!(n-x)!} \right) \cdot \pi^x \cdot (1 - \pi)^{n-x}$$

Where:

x = the number of successes 0, 1, 2, etc

n = number of trials or sample size

π = probability of success on each trial

Steps

1. Find the probability (π) of a success in each trial.
2. Find the number of trials (n).
3. Decide on the number of successes (x) for which you want to determine the probability.
4. Substitute the values into the formula.

Example 9.2

Suppose the probability is 0.2 that any given avocado will show measurable damage when the temperature falls to 15°C. If the temperature does drop to 15°C, construct the binomial distribution for a sample of five avocados.

$$\pi = 0.2$$

$$n = 5$$

$x = 0, 1, 2, 3, 4, 5$ (Damage can occur in none of the five, one of the five ... up to all five).

x	$P(x)$
0	0.3277
1	0.4096
2	0.2048
3	0.0512
4	0.0064
5	0.0003
	1.0000

$$P(x=0) = \frac{5!}{0!(5-0)!} \cdot 0.2^0 \cdot (1-0.2)^{5-0} = 0.3277$$

$$P(x=1) = \frac{5!}{1!(5-1)!} \cdot 0.2^1 \cdot (1-0.2)^{5-1} = 0.4096$$

$$P(x=2) = \frac{5!}{2!(5-2)!} \cdot 0.2^2 \cdot (1-0.2)^{5-2} = 0.2048$$

$$P(x=3) = \frac{5!}{3!(5-3)!} \cdot 0.2^3 \cdot (1-0.2)^{5-3} = 0.0512$$

$$P(x=4) = \frac{5!}{4!(5-4)!} \cdot 0.2^4 \cdot (1-0.2)^{5-4} = 0.0064$$

$$P(x=5) = \frac{5!}{5!(5-5)!} \cdot 0.2^5 \cdot (1-0.2)^{5-5} = 0.0003$$

- The probability that none are damaged: $P(x=0) = 0.3277$
- The probability that all five are damaged: $P(x=5) = 0.0003$
- The probability that less than two are damaged:
Calculate $P(x=0)$ and $P(x=1)$ and add the answers.
 $P(x=0 \text{ or } 1) = [P(x=0) + P(x=1)] = 0.3277 + 0.4096 = 0.7373$

Note: A probability is the sum of the probabilities of all the possibilities that will give you the answer.

- The probability that at least two will be damaged:
 $P(x \geq 2) = P(x = 2, 3, 4 \text{ or } 5) = 1 - P(x = 0 \text{ or } 1)$
 $= 1 - (0.3277 + 0.4096)$
 $= 0.2627$

Note: You can apply the complement rule or you can calculate $P(x = 2, 3, 4 \text{ or } 5)$. It will give you the same answer. Use the complement rule if it will give you a short-cut method to your answer.

Activity 9.1

A shoe store's records show that 30% of customers making a purchase use a credit card to make payment. This morning seven customers purchased shoes from the store. What is the probability that:

1. three customers will pay using a credit card?
2. at least two customers will pay using a credit card?
3. more than five customers will pay using a credit card?
4. exactly three customers will not pay using a credit card?

9.1.2 Poisson distribution

The Poisson distribution is a discrete distribution and is useful for calculating the probability that a certain number of successes ($x = 0, 1, 2, \dots$) will occur over a specific interval of time, area or other measurement. There is no specific upper limit to the count (n is unknown) although a finite count is expected.

Characteristics

1. The number of successes that occur in one interval is independent of the number of successes that occur in any other interval.
2. The probability that a success will occur in an interval is the same for all intervals of equal size and is proportional to the size of the interval.
3. x is the count of the number of successes that occur in a given interval of time or other measurement and may take on any value from 0 to infinity.
4. If x is a Poisson random variable, the probability distribution of x is given by:

$$P(x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}$$

Where:

$x = 0, 1, 2, \dots$

λ (pronounced as *lambda*) = number of successes in the given unit of measurement

e = the base of natural logarithms (use the e^x -key on the calculator)

Example 9.3

If a company receives an average of three calls per five-minute period of the working day, what is the probability that:

1. no calls will be received during a randomly selected five minutes?

$\lambda = 3$ per five minutes

$$P(x=0) = \frac{3^0 \cdot e^{-3}}{0!} = 0.0498$$

2. five calls will be received during the next 10 minutes?

$\lambda = 3$ for five minutes therefore $\lambda = 6$ for 10 minutes

$$P(x=5) = \frac{6^5 \cdot e^{-6}}{5!} = 0.1606$$

3. at least two calls will be received during the next 2.5 minutes

$\lambda = 3$ for five minutes therefore $\lambda = 1.5$ for 2.5 minutes

$$P(x=0) = \frac{1.5^0 \cdot e^{-1.5}}{0!} = 0.2231$$

$$P(x=1) = \frac{1.5^1 \cdot e^{-1.5}}{1!} = 0.3347$$

$$\begin{aligned} P(x \geq 2) &= P(x = 2, 3, 4, \dots) = 1 - P(x = 0, 1) \\ &= 1 - (0.2231 + 0.3347) \\ &= 0.4422 \end{aligned}$$

Note: There is no n value available and therefore no upper limit to the x -counts. If you are required to do a probability involving $>$ or \geq , you have to apply the complement rule.

Activity 9.2

A tollgate operator has observed that cars arrived randomly at an average of 360 cars per hour. Calculate the probability that:

1. only two cars will arrive during a specified one-minute period
2. at least three cars will arrive during a specified two-minute period.

9.2 Probability distributions for continuous random variables

A continuous distribution is a distribution in which the x -variable may assume any value within a given range or interval. The most widely used continuous probability distribution is the normal distribution.

9.2.1 The normal distribution

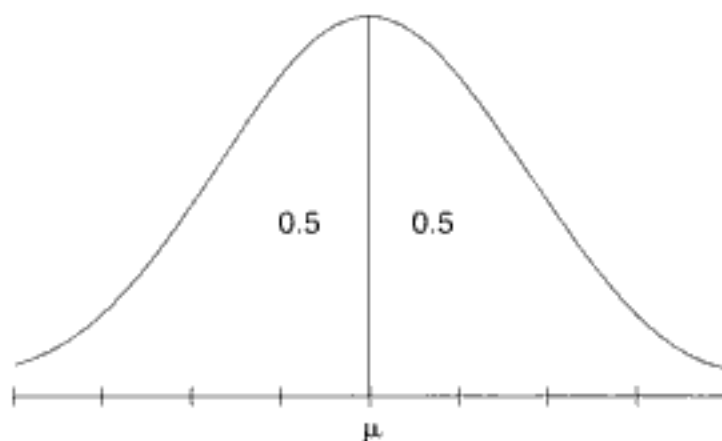
In general, many things are distributed with the characteristics of what we call normal. That is, there are lots of events or occurrences in the middle of the distribution, but relatively few on each end. For example, there are relatively few tall people and relatively few short people, but lots of people in the middle. That means the chance (or probability) that a person will be average in height, or more or less in the middle of the distribution, will be much better than the chance that any one person will either be very tall or very short.

Those events that tend to occur in the middle of the normal curve have a higher probability of occurring than the ones in the extreme.

The normal distribution plays a very important role in statistical inference.

Characteristics

- The graph for a continuous random x -variable is a smooth curve.
- The curve is unimodal (single mode).



Hidden page

How to read an area from the normal table using a z-score

Steps

- The first column in the table gives the z-score to the first decimal place, and the top row gives the second decimal for a z-score.
- For example, to find the area of a z-score of 0.23:
 - find 0.2 in the first column
 - go across with this row up to the column headed with the second decimal of 0.03
 - where the corresponding row and column intersects, the area is 0.0910
 - this is the area between a z-score of 0 and 0.23 and is denoted as $P(0 \leq z \leq 0.23)$
 - a z-score of 1.02 or $P(0 \leq z \leq 1.02)$ will correspond to an area of 0.3461.
- This table always gives the area between the mean and the required z-score.

Sample of the standard normal table from appendix 1.

z	0.00	0.01	0.02	0.03
0.0	0.0000	0.0040	0.0080	0.0120
0.1	0.0398	0.0438	0.0478	0.0517
0.2	0.0793	0.0832	0.0871	0.0910
0.3	0.1179	0.1217	0.1255	0.1293
↓	↓	↓	↓	↓
1.0	0.3413	0.3418	0.3461	0.3485

9.2.2 Different areas under the normal curve

Steps

- Draw the normal curve and indicate the mean in the middle.
- Find the value of the x-variable on the x-axis and shade the desired area.
- Calculate the z-score using the z-formula.
- Use the standard normal table and the *absolute value* of the calculated z-score to find corresponding areas and probabilities.

Area between μ and any x -value

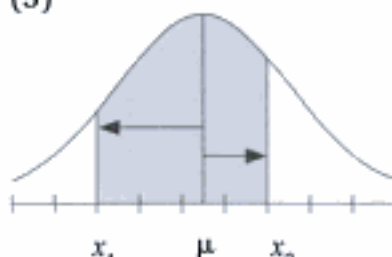
(1)



(2)



(3)

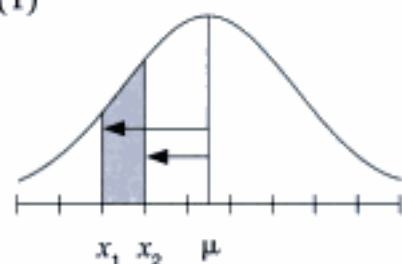


Steps

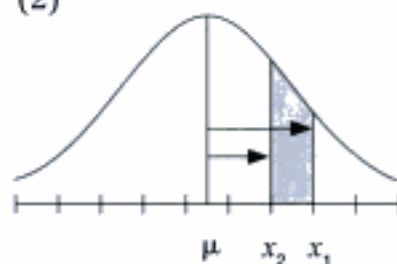
1. Sketch (1): Calculate the z -score and look up the value in the normal table.
2. Sketch (2): Calculate the z -score. The answer for z is negative, therefore use the absolute value of z and look up the value in the normal table.
3. Sketch (3): If the area to be determined falls on **both sides** of the mean:
 - calculate the z -score for the area between μ and the x -value to the right of μ
 - determine the z -score for the area between μ and the x -value to the left of μ
 - look up the areas for the two z -scores from the normal table
 - add the two areas.

Area between two x -values on the same side of the mean

(1)



(2)

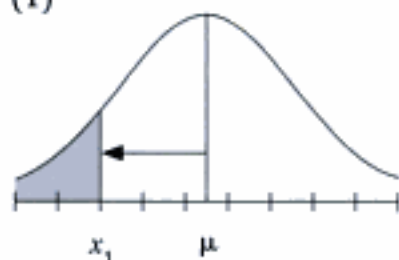


Steps

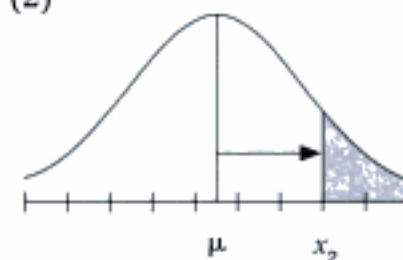
1. Calculate the z -score for the area between μ and the larger x -value.
2. Calculate the z -score for the area between μ and the smaller x -value.
3. Look up the two z -scores in the normal table to obtain the two areas.
4. Subtract the smaller area from the larger area.

Area in any tail

(1)



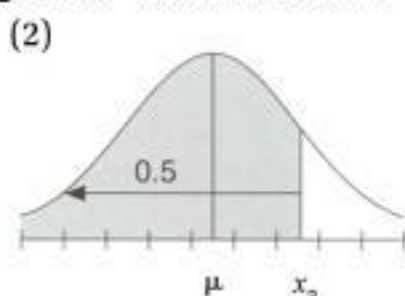
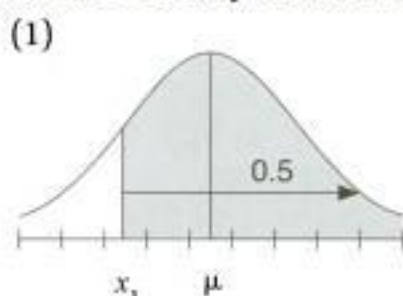
(2)



Steps

1. Calculate the z -score for the area between μ and the x -value.
2. Use the absolute z -score to look up the area in the normal table.
3. Subtract the area from 0.5 (the area from the mean to the end of the distribution is 0.5).

Area to the right of any x -value, where x is less than the mean and the area to the left of any x -value, where x is greater than the mean



Steps

1. Calculate the z -score for the area between μ and the x -value
2. Look up the z -score to get the area.
3. Add 0.5 to the area.

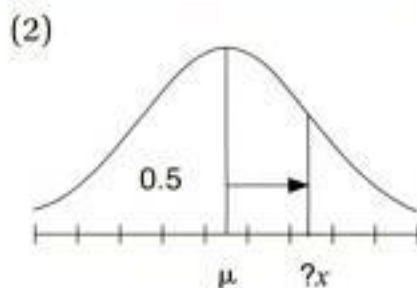
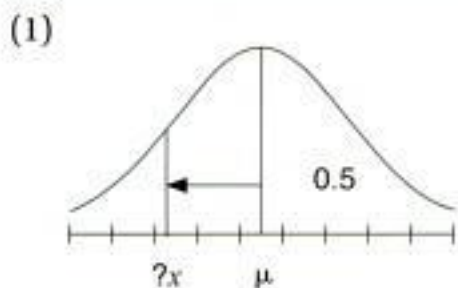
Find the z -score for a given area

Steps

If a probability or area is given and you are required to determine an unknown x -value:

1. Calculate the area between μ and the unknown x -value.
2. If the known area falls in the tail of the curve, subtract the tail area from 0.5.
3. Compare this area with the areas in the body of the normal table.
4. If the exact area is not listed, use the closest value.
5. Read the z -score from the first column and top row.
6. Use this z -score in the z -formula to obtain the unknown x -value.

Note: If the area falls to the left of μ , the z is negative, and if the area falls to the right of μ , the z is positive.



Hidden page

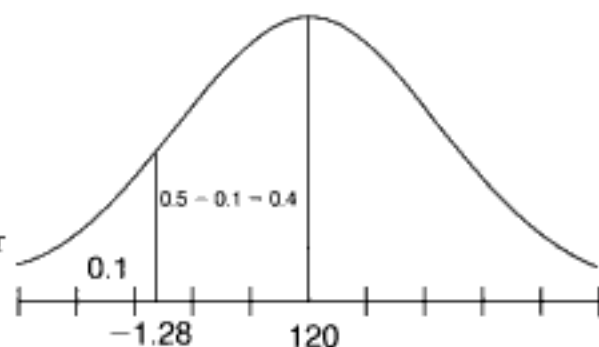
5. The 10% of the employees who complete the task within the shortest time are to be given advance training. What task times qualify individuals for such training?
- The fastest 10% will fall in the left-hand tail of the distribution because those times will be the shortest.
 - The area between the mean and the 10% in the tail is: $0.5 - 0.1 = 0.4$
 - Look up an area of 0.40 in the normal table and obtain the z -value. The area you are interested in falls on the left of μ , resulting in a negative z -score.

$$z = \frac{x - \mu}{\sigma}$$

$$-1.28 = \frac{x - 120}{20}$$

$$\therefore x = 94.4$$

This means that the applicants who complete that task in 94.4 seconds or less, will qualify for advance training.



Activity 9.3

The lifetimes of a certain kind of battery have a mean of 300 hours and a standard deviation of 35 hours. Assume that the lifetimes, measured to the nearest hour, follow a normal distribution, and determine:

1. the percentage of batteries that have a lifetime of more than 320 hours
2. the value above which the best 30% of the batteries lie
3. the proportion of batteries that have a lifetime from 250 to 350 hours
4. the proportion of batteries with a lifetime between 250 and 280 hours
5. the maximum lifetime below which the weakest 20% of the batteries will fall
6. the minimum lifetime above which the 15% of the batteries with the longest life will fall.

TEST YOURSELF 9

1. A textile firm has found from experience that only 20% of the people applying for a certain stitching-machine job are qualified for the work:
 - a) Construct the probability distribution for this investigation if five persons are interviewed to find qualified persons.
 - b) What is the probability that at least two are qualified for the job?
2. Testing for the presence of antibodies in blood for HIV, the virus that causes Aids, gives a positive result with probability of about 0.004 when a person who is free of HIV antibodies is tested. A clinic tests three people who are all free of HIV antibodies:
 - a) Construct the probability distribution for this investigation.
 - b) What is the probability that you will get one false-positive result?
 - c) What is the probability that you will get more than one false-positive result?
3. You read that one out of four eggs contains salmonella bacteria. If you use six eggs in your chocolate cake, what is the probability that:
 - a) one of the eggs contains salmonella?
 - b) at most two of the eggs contain salmonella?

4. If 40% of all patients have medical aid, what is the probability that in a sample of 10 patients:
 - a) exactly four will have medical aid?
 - b) at least four will have medical aid?
 - c) at most four will have medical aid?
5. Shortly after being put into service, some buses of a certain type develop cracks on the underside of the mainframe. A particular city has 20 buses of this type, eight of which have cracks. If five buses are randomly selected for inspection, determine the probability of finding:
 - a) exactly two buses with cracks
 - b) at most two buses with cracks.
6. About 15% of the population is left-handed. Fifteen individuals are randomly selected. What is the probability that:
 - a) three or fewer are left-handed?
 - b) one or more are right-handed?
7. According to the National Environmental Program, air pollution standards for particulate matter are exceeded an average of 5.6 days in every three-week period. What is the probability that the standard is:
 - a) not exceeded on any day during a three-week period?
 - b) exceeded two days or more of a two-week period?
8. S.A Flawless Steel Co. produces stainless steel plating that has an occasional defect once every 10 m². What is the probability that:
 - a) a square metre of stainless steel plating will have no defects?
 - b) two square metres of stainless steel plating will have exactly one defect?
 - c) five square metres of stainless steel plating will have two or more defects?
9. In a local bakery, the manager monitored the customers' arrivals for Saturdays. She estimated the average number of customer arrivals per 10-minute period to be 6.2. What is the probability of:
 - a) 10 customers entering during a half-hour interval?
 - b) six customers entering during a five-minute interval?
 - c) 15 customers entering during an hour?
 - d) at least one customer entering during a 10-minute interval?
10. A welding machine breaks down occasionally as a result of a particular part that wears out, and these breakdowns occur on average four times per eight hour day. Find the probability that:
 - a) no breakdown will occur during a given day
 - b) at most two breakdowns will occur during the first hour of the day.
11. The photocopier repair department of Papermate receives an average of two calls for service per hour. What is the probability of:
 - a) receiving no service calls per hour?
 - b) receiving exactly two service calls in two hours?
 - c) receiving more than three service calls in 1.5 hour?
 - d) receiving no service calls in the next half hour?
12. An environmental study has shown that the daily average noise level on a busy street follows a normal distribution with a mean of 37 decibels and standard deviation of six decibels:
 - a) What is the probability that the noise level exceeds 46 decibels?

- b) What decibel range contains the middle 95% of the distribution?
 c) What is the probability that the noise level will be between 20 and 30 decibels?
13. Accurate labelling of packaged meat is difficult because of weight decrease due to moisture loss. Suppose that moisture loss for a package of chicken breasts is normally distributed with a mean value of 4% and standard deviation of 1%. What is the probability that moisture loss is:
- a) between 3% and 5%
 b) at most 4%
 c) at least 7%?
 and
 d) 90% of all packages have moisture losses below what number?
14. Manufactured items are sold in boxes that are stated to contain a mass of at least 40 kg. The actual mass in a box varies with a mean of 41.2 kg and a standard deviation of 0.8 kg:
- a) Calculate the proportion of boxes whose mass is between 40 kg and 42 kg.
 b) Calculate the mass below which 20% of the lightest boxes fall.
 c) All boxes containing less than 40 kg are scrapped at a cost of R100 per box. Calculate the scrapping cost associated with the packing of 50 boxes.
 d) To what mean mass should the box contents be adjusted, with the standard deviation unchanged, if only 1% of the boxes are to be scrapped?
15. The production foreman of the Oros Fruit Company estimates that the average sales of oranges is 4 700 and the standard deviation 500 oranges. Calculate the probability that sales will be:
- a) more than 5 500 oranges?
 b) more than 4 500 oranges?
 c) less than 4 900 oranges?
 d) between 4 500 and 4 900 oranges?
 e) between 4 900 and 5 500 oranges?
16. Birth weights are normally distributed with a mean of 3 579 g and a standard deviation of 500 g. What is the cut-off point for the lightest 2% of the babies?
17. The Faber Co. produces a pencil called Ultra-Light. Sales follow a normal distribution with a mean of 457 000 pencils each year. Furthermore, 90% of the time sales have been between 460 000 and 454 000 pencils. Estimate the standard deviation of this distribution.
18. The average number of calories in a 50 g chocolate bar is 225. If the distribution of calories is approximately normal with a standard deviation of 10, find the probability that a randomly selected chocolate bar will have between 200 and 220 calories.
19. The thickness of bolts (mm) manufactured by a certain process follows a normal distribution with a mean of 10 mm and a standard deviation of 1 mm:
- a) What proportion of the bolts in the long-run are at most 11 mm?
 b) What proportion of the bolts will have thickness values between 7.5 mm and 12.5 mm?
 c) What proportion of bolts will have thicknesses that exceed 11.5 mm?
20. The amount of distilled water dispensed by a certain machine has a normal distribution with a mean of 64 l and a standard deviation of 0.78 l. What container size will ensure that overflow occurs only 0.5% of the time?

UNIT 10

Statistical inference: estimation

The objective of most statistical studies is inference. Inferential methods presented in the following two units use information contained in a sample to reach conclusions about one or more characteristics of the population from which the sample was drawn. There are two general procedures for making inferences about populations: estimation and hypothesis testing.

After completion of this unit you will be able to:

- describe the purpose of sampling distributions
- find the confidence interval for the mean
- find the confidence interval for the proportion
- determine the size of the sample required in order to make estimates to a specified degree of accuracy.

10.1 Statistics and parameters

Calculations based on a sample are known as sample statistics. The values calculated from population information are referred to as population parameters. To distinguish between the two, Greek letters will be used to refer to population parameters and Roman letters will refer to sample statistics.

	Statistic	Parameter
Mean	\bar{x}	μ
Standard deviation	s	σ
Proportion	p	π
Size	n	N

10.2 Sampling distribution of the means

The sampling distribution theorem forms the foundation of inferential statistics. The theorem states that if you draw an infinite number of different samples of the same size (n) from a population, you can calculate the mean of each sample and these means will

most probably differ. If you list all these possible samples together with their means, it is known as the **sampling distribution of the means**.

Properties of the sampling distribution of the means

- If you calculate the mean of all the different sample means it is equal to the population mean: $\mu_{\bar{x}} = \mu$
- The differences between the means are known as the variability in the sampling distribution of the means and can be measured by the standard error of the mean: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. The larger the sample size, the smaller the standard error of the mean and the better the estimate of the population mean because of the lesser dispersion.
- In most cases the population standard deviation (σ) is unknown and the sample standard deviation (s) is used as an approximation to the population standard deviation (σ). The standard error of the mean then becomes: $s_{\bar{x}} = \frac{s}{\sqrt{n}}$
- If the population is normally distributed **with a known population** σ , the sample distribution of the mean is also normal, *regardless of the sample size*.
- The *central limit theorem* states that if the population from which the sample is drawn is not normal, the distribution of the sample means will become more and more normal as the sample size increases. A sample of $n \geq 30$ is considered by most as being large enough to assume a normal distribution.
- In a population that is normal or close to normal **with an unknown population** σ , the distribution of sample means is referred to as the student's *t*-distribution. The sample standard deviation (s) is used as an estimate of σ .

The *t*-distributions are more dispersed than the normal distribution and are distinguished by a positive whole number called degrees of freedom: **$df = n - 1$**

Degrees of freedom (*df*) is defined as one less than the sample size ($n - 1$) and it represents the number of observations that are 'free to vary' around the mean of the sample.

There is a different *t*-distribution for each sample size. But, as the sample size gets larger the *t*-distribution becomes more and more normal. Once the number of degrees of freedom exceeds 30, the *t*-distribution is so close to the normal distribution that we can use the normal distribution to approximate the *t*. That means that the *t*-distribution will only be used for samples with sizes of less than 30.

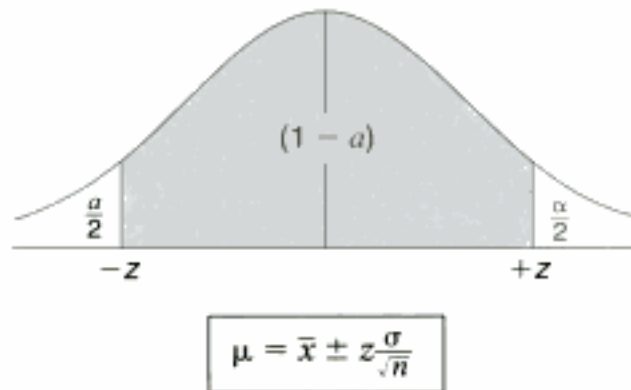
- Sometimes the parameter of interest is not the arithmetic mean, but a proportion (π). The sample proportion (p) is used to estimate the population proportion (π). Sample proportions will vary from sample to sample from a given population in the same way that sample means vary.

The sample distribution of proportions is assumed to be normal if $np \geq 5$ and $n(1 - p) \geq 5$. The mean is: $\mu_p = \pi$

The standard error is: $\sigma_p = \sqrt{\frac{(\pi(1 - \pi))}{n}}$

If π is unknown, p can substitute π : $\sigma_p = \sqrt{\frac{p(1 - p)}{n}}$

Hidden page

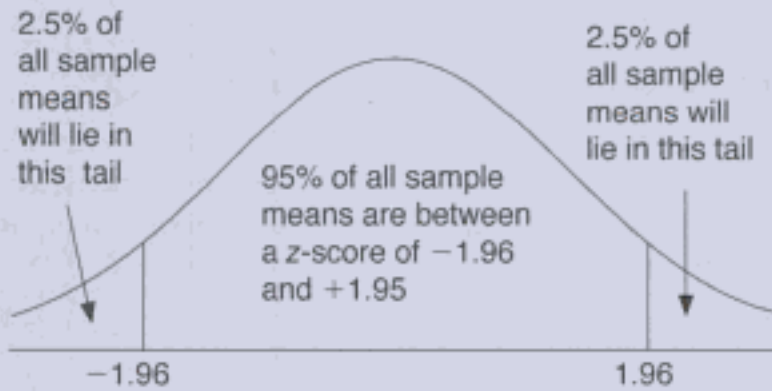


- \bar{x} is the point estimator for μ .
- z is the critical value associated with the chosen level of confidence.
- $\frac{\sigma}{\sqrt{n}}$ is the standard error.
- $z \cdot \frac{\sigma}{\sqrt{n}}$ is known as the margin of error (E).
- The lower boundary of the interval is $\bar{x} - z \cdot \frac{\sigma}{\sqrt{n}}$
- The upper boundary of the interval is $\bar{x} + z \cdot \frac{\sigma}{\sqrt{n}}$

Calculating a confidence interval

Steps

1. Collect a sample of an adequate size (n).
2. Compute the sample mean (\bar{x}) and standard deviation (s) or proportion (p).
3. Determine the type of sampling distribution:
 - normal (z) if population is normally distributed with known σ
 - normal (z) via the central limit theorem with ($n \geq 30$) and known σ
 - normal (z) via the central limit theorem ($n \geq 30$) if σ is unknown
 - student t -distribution for $n < 30$ and unknown σ
 - normal (z) if dealing with proportions with $n\pi \geq 5$ and $n(1 - \pi) \geq 5$
4. Identify the level of confidence ($1 - \alpha$): a 95% level implies that if 100 different confidence intervals are constructed, each based on a different sample from the same population, we expect 95 of the intervals to include the parameter and five not to include the parameter. We are capturing the middle 95% between the two critical values and 2.5% in each tail.
5. Find the critical value z or t that corresponds to the level of confidence by making use of the *appropriate* table (normal z -table or student t -table) and the level of confidence. A critical value is the cut-off point between the sample statistics that are likely to occur and those that are unlikely to occur.



For a confidence level of 95%: the standard normal table is used to determine a value z such that a central area of 0.95 falls between $-z = -1.96$ and $+z = 1.96$.

To identify the critical z -value for a 95% confidence level we know that 95% or 0.95 covers the middle area of the curve. Do not look up 0.95 in the body of the normal table, because the normal table contains probabilities only for one half of the normal curve. Divide 0.95 by two to obtain the area to the left or right of the mean [$0.95 \div 2 = 0.475$]. Look up 0.475 or the closest to this area in the body of the table to find the corresponding z as ± 1.96 (\pm because the area to the left of the mean will result in a $-z$ and to the right of the mean a $+z$).

6. Find the margin of error (E) which is the critical value multiplied by the standard error of estimate: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ or $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$
7. Calculate the upper and lower confidence limits by making use of the appropriate formula:
 - $\mu = \bar{x} \pm z \frac{\sigma}{\sqrt{n}}$
 - $\mu = \bar{x} \pm t \frac{\sigma}{\sqrt{n}}$
 - $\pi = p \pm z \sqrt{\frac{p(1-p)}{n}}$
8. Briefly state the meaning of your confidence interval: a confidence interval indicates that, if we obtain many samples of size n from the population whose mean, μ , is unknown, then approximately $(1 - \alpha) \cdot 100\%$ of the intervals will contain μ or π .

Activity 10.1

Identify the critical z -values associated with a confidence level of: 90%, 98% and 99%.

10.3.3 Confidence interval estimate for the population mean (μ) for data obtained from a population that is normally distributed or from large samples ($n \geq 30$)

The central limit theorem states that if n is large ($n \geq 30$), the sampling distribution of the mean will be approximately normal. It does not matter whether σ is known or unknown or if the distribution is normal or not. If σ is unknown, substitute σ with the sample standard deviation s .

Example 10.2

The Pappi Paper Company wanted to estimate the average time required for a new machine to produce a ream of paper. A sample of 36 reams required an average production time of 1.5 minutes for each ream. The population standard deviation was 0.30 min and the confidence level was 95%.

1. The population distribution is not known to be normal but via the central limit theorem we assume a normal distribution.
2. To obtain the z -value from the normal table, divide the confidence level by two: ($0.95 \div 2 = 0.475$) and look up the area in the body of the normal table. An area of 0.475 corresponds to $z = \pm 1.96$
3. Use the sample standard deviation (s) to estimate the population σ .
4. Use the normal distribution formula to calculate the interval boundaries.

$$\begin{aligned}\mu &= \bar{x} \pm z \frac{\sigma}{\sqrt{n}} \\ &= 1.5 \pm 1.96 \cdot \frac{0.3}{\sqrt{36}} \\ &= 1.5 \pm 0.098 \quad [1.5 - 0.098 = 1.402 \text{ and } 1.5 + 0.098 = 1.598] \\ \therefore 1.402 &\leq \mu \leq 1.598\end{aligned}$$

5. Based on the sample data, the Pappi Paper Company can be 95% confident that the average time required for a new machine to produce a ream of paper lies between 1.402 and 1.598 minutes.

Activity 10.2

In 36 randomly selected seawater samples, the mean sodium chloride concentration was 23 cm³ per m³ and the standard deviation was 6.7 cm³ per m³. Construct a 98% confidence interval estimate for the mean sodium chloride concentration.

10.3.4 Confidence interval estimate for the population mean using small samples ($n < 30$) with σ unknown: t-distribution

Example 10.3

The number of home fires that were started by candles in low-cost housing areas was recorded for a sample of seven years. The mean number of fires was 7 046 with a standard deviation of 1 605. Calculate the 99% confidence interval for the average number of home fires started by candles.

- Find the critical values for 99% confidence and $n = 7$ using the t -table from appendix 2:
 $df = n - 1 = 7 - 1 = 6$
 Use the t -table to look up the critical values. The top row of the table indicates a one-tail test or a two-tail test at a specified significance level (α). All confidence level interval estimates are two-tail tests. If the level of confidence is 0.99, the α -value is: $(1 - 0.99) = 0.01$. Choose the 0.01 column under the two-tail row and go down in this column to where it corresponds with the desired degrees of freedom (df), which is 6. The df column is the first column in the table. This t -table value is 3.707.
- $\mu = \bar{x} \pm t \frac{s}{\sqrt{n}}$
 $= 7\,046 \pm 3.707 \frac{1\,605}{\sqrt{7}}$
 $= 7\,046 \pm 2248.8$
 $\therefore 4\,797.2 \leq \mu \leq 9\,294.8$
- Based on the sample data, we can be 99% confident that the mean number of home fires per year will be between 4 797 and 9 295.

Activity 10.3

The time taken to complete the same task (in minutes) was recorded for nine participants in a training exercise as follows:

8 7 8 9 7 7 9 10 9

Construct a 95% confidence interval for the average time taken to complete the task.

10.3.5 Confidence interval estimate for the population proportion (π)

Steps

- Identify the sample statistics n and x (or p).
- Find the point estimate if not given: $p = \frac{x}{n}$
- Verify that the sampling distribution of p can be approximated by the normal distribution.
- Find the critical value z that corresponds to the given level of confidence.
- Use the formula to calculate the margin of error and the confidence interval boundaries.

Example 10.4

A survey found that out of 200 workers, 168 said they were interrupted three or more times an hour by phone calls. Find the 90% confidence interval of the population proportion of workers who are interrupted three or more times an hour.

Hidden page

Hidden page

TEST YOURSELF 10

1. A survey of 100 customers passing through the check-out line of a supermarket revealed a mean check-out time of 190 seconds and a standard deviation of 60 seconds. Construct a 90% confidence interval for the true mean checkout time.
2. The calories per 125 g serving of ice cream are recorded for a sample of 16 popular chocolate ice cream brands. The mean calories were found to be 190 calories with a standard deviation of 40 calories. Construct a 98% confidence interval estimate for the mean calorie content of chocolate ice cream.
3. Noise levels at various hospitals in Gauteng are measured. The mean of the noise level in 84 corridors was 61.2 decibels, and the standard deviation was 7.9. Find the 99% confidence interval of the true noise level mean.
4. The Department of Health has been concerned about lead levels in South African wines. In a previous testing of 40 wine specimens, lead levels of 600 parts per billion were recorded with a standard deviation of 50 parts per billion. Estimate the true lead level for the wine using a 90% level of confidence.
5. The tear strength of a particular paper product is known to be normally distributed. If a random sample of nine rolls yielded a mean tear strength of 225 kg/m² with a standard deviation of 15 kg/m², construct a 90% confidence interval estimate for the average tear strength.
6. A sample of 12 households in Johannesburg showed a mean of R50 expenditure per day with a standard deviation of R20. If household expenditure follows a normal distribution, construct a 99% confidence interval estimate for daily household expenditure for all households in Johannesburg.
7. Not all the town's electricity accounts have been paid on time. The town clerk takes a random sample of 16 from the outstanding accounts file and finds the mean amount owed to be R230 with a standard deviation of R40. If there are 100 outstanding accounts, find a 99% confidence interval estimate for the mean amount in outstanding bills.
8. Fat content (in %) for 10 randomly selected hot dogs are listed below:

25.2 21.3 22.8 17.0 29.8 21.0 25.5 16.0 20.9 19.5

Construct a 90% confidence interval for the true mean fat percentage of hot dogs.

9. For a group of 10 men subjected to a stress situation, the mean number of heart beats per minute was 126 and the standard deviation was four. Find the 95% confidence interval for the true mean.
10. A sample of 150 Gauteng residents found 65% to be in favour of fluoridation of drinking water. Construct a 95% confidence interval for the true proportion of Gauteng residents who favour fluoridation.
11. A sample of 200 manufactured items contains 40 defectives. Construct a 90% confidence interval for the true percentage of defectives.
12. You want to determine with 98% confidence the proportion of adults age 20 to 29 that have high blood pressure. If a sample of 60 adults in this age group showed 4% with high blood pressure, what will your confidence boundaries be?

13. A cereal manufacturer has recently redesigned its product packaging. In a random sample of 1 000 households prior to the change, the manufacturer found that 220 were satisfied with the packaging. Estimate with a 98% confidence the proportion of customers that were satisfied with the old container.
 14. In a survey of 80 adults, it was found that 72 ate the recommended amount of fruits and vegetables each day. Construct a 99% confidence interval for the proportion of this population that follows these recommendations.
 15. In a study to determine the proportion of adult males who have hypertension, what sample size would be needed for the estimate to be within 3% at a 95% confidence level? A previous study showed 9% of adult males had hypertension.
 16. To study the proportion of residents who live in neighbourhoods with acceptable levels of carbon monoxide, what sample size would be needed for the estimate to be within 1.5% of the proportion with 90% confidence? A previous study showed that 90% of the residents live in neighbourhoods with acceptable levels of carbon monoxide.
 17. Motorola wishes to estimate the mean talk time for its V505 camera phone before the battery must be recharged. If $s = 31$ minutes, how many phones would Motorola need to test to estimate the mean talk time within five minutes with 95% confidence?
 18. The lives of light bulbs are normally distributed with a known standard deviation of 60 hours. If you want to estimate the sample mean within three hours of the true population mean at a 95% confidence, what sample size is needed?
 19. A meat packer is investigating the marked mass shown on Vienna sausages. A pilot study showed a mean mass of 11.8 kg per pack and a standard deviation of 0.7 kg. How many packs should be sampled in order to be 98% confident that the sample mean will differ by at most 0.2 kg?
 20. A study is planned to determine the average annual family medical expenses of government employees. A previously known standard deviation is R400 and the analyst wants to be 95% confident that the sample average is within R50 of the true family expenses. How large a sample is necessary?
 21. A publishing company wants to estimate the proportion of its customers that would purchase TV programme guides. A 95% confidence is required that the estimate is correct within 5% of the true proportion. If past experience in other areas indicates that 30% of the customers will purchase the programme guide, what is the sample size needed?
-

UNIT 11

Hypothesis testing

In this unit you will continue your study of inferential statistics and will learn how to test a claim or hypothesis about a population parameter.

After completion of this unit you will be able to:

- explain the reasoning behind hypothesis testing
- explain the steps in the hypothesis-testing procedure
- distinguish between a one-tailed and two-tailed test
- conduct tests of hypotheses concerning values of the following parameters:
 - population mean; large and small samples
 - population proportion
- conduct chi-square tests.

A hypothesis is a claim or statement about a population characteristic.

Hypothesis testing is a decision-making process to determine whether enough statistical evidence exists to enable us to conclude that a belief or hypothesis about a population parameter is reasonable.

To make this decision, it is necessary to decide whether the difference that exists between the hypothesised population parameter and the sample result is significant and therefore not supportive of the hypothesis, or whether the difference is a chance difference and therefore supportive of the claim.

In the sampling process, any one of the samples in the sampling distribution might be selected. Most of the time this sample mean would not equal the population mean. Such a difference is due to the sampling process and is known as a **chance difference**. The difference is not large enough to cause concern.

Results are **statistically significant** if the difference between the sample result and the statement made in the null hypothesis is unlikely to occur due to chance alone. It indicates that the sample came from a population with a mean other than the hypothesised mean.

11.1 Steps to follow in a single sample hypothesis test

Understand the problem

1. Set up the null and alternative hypothesis.

Define the test procedure and decision rule

2. Select the significance level (α) for the test.
3. Determine the type of sampling distribution (z or t)
4. Determine the critical value(s) and corresponding rejection region.

Collect and analyse the data

5. Collect the data, and calculate the test statistic.

Draw conclusions and make recommendations

6. Make the statistical decision by comparing the test statistic with the rejection region.
7. Interpret the statistical decision.

11.1.1 Stating hypotheses

To test a population parameter, you should identify a pair of hypotheses; one that represents the claim (H_0) and the other its complement (H_A).

H_0 : population characteristic (μ) = hypothesised value

- A null hypothesis, denoted by H_0 , is a statement of equality or no difference. The 'null' implies that there has been no change or no difference in the value of the parameter.
- The null hypotheses is assumed true until evidence indicates otherwise.

H_A : state the alternative hypothesis

- The H_A (or test hypothesis) states what the case will be if the null hypothesis is not true.
- The value of the parameter appearing in H_A must be identical to the one used in H_0 .
- Either hypothesis – the H_0 or the H_A – may represent the original claim.

Different ways to set up the hypothesis:

- A test in which we want to find out whether a population parameter (μ or π) *has changed* (\neq), regardless of the direction of change, is referred to as a **two-tailed test**.

H_0 : population characteristic = hypothesised value (will remain unchanged)

H_A : population characteristic \neq hypothesised value (will be different)

- If we wish to determine whether the sample came from a population that has a parameter (μ or π) *less than or more than* a hypothesised value, the attention is focused on the direction of change, and the test is referred to as **left-tailed** or **right-tailed**.

H_0 : population characteristic = hypothesised value

H_A : population characteristic < hypothesised value

or

H_0 : population characteristic = hypothesised value

H_A : population characteristic > hypothesised value

Some common phrases that indicate the direction of test

> (right-tailed)	< (left-tailed)	= or \neq (two-tailed)
Is greater than Is above Is higher than Is longer than Is bigger than Is increased Is at least Is not less than An incline	Is less than Is below Is lower than Is shorter than Is smaller than Is decreased or reduced Is at most Is not more than A decline	Is equal to or not equal to Is the same as or different from It has not changed or it has changed from

11.1.2 Select a level of significance (α) to be used.

When you perform a hypothesis test, you can make one of two decisions: reject H_0 or do not reject H_0 . Because this decision is based on a sample and not on the population, there is always a possibility that the decision could be wrong. The probability associated with this uncertainty is your level of significance. Just as we place a level of confidence in the construction of an interval, we can determine the probability of making errors. The significance level is chosen by the researcher before the sample data are collected.

- α = type I error and occurs if H_0 is rejected when it is true.
- Type II error occurs if H_0 is not rejected when it is false.

For example, when $\alpha = 0.10$, there is a 10% chance of rejecting a true H_0 . You can decrease the probability of rejecting H_0 when it is actually true by lowering the significance level.

The significance level is the maximum probability of making a type I error and is denoted by α .

- The purpose of the level of significance is to provide a probability basis for deciding whether an observed difference between a sample statistic and a hypothesised parameter is a chance difference or a statistically significant difference, since α is the probability that the test statistic will fall in the rejection area.
- Usually tests are performed at an α -value of 0.01, 0.02, 0.05 or 0.10.

11.1.3 Determine the type of sampling distribution

This step will enable you to know whether to use the normal z -table or the t -distribution table in determining the rejection area.

Use the normal z -distribution:

- when the distribution is approximately normal with a known σ
- when the sample size $n \geq 30$ with an unknown or known σ .

Use the t -distribution:

- when σ is unknown and the sample size $n < 30$. If $n \geq 30$, the distribution approximates the normal curve and you use the normal z -table.

11.1.4 Determine the critical value(s) and identify the rejection region

- The **critical value** represents the maximum number of standard deviations the sample mean or proportion can differ from the hypothesised value before H_0 is rejected.
- The critical value separates the area under the curve into two regions – the non-rejection region and the rejection region.
- The **rejection region** (or decision rule) is a range of values such that, if the test statistic falls into that range, we reject H_0 .

Steps

1. Specify the level of significance (α).
2. Decide whether the test is two-tailed, left-tailed or right-tailed.

The H_A is the indicator whether to do a *one-tailed* test or a *two-tailed* test.

- If $H_A: \mu \neq$ hypothesised value: two-tailed test
 - If $H_A: \mu <$ hypothesised value: left-tailed test
 - If $H_A: \mu >$ hypothesised value: right-tailed test
3. Find the critical value(s). One or two critical values are established on the horizontal axis of the distribution, which serve as cut-off points between the non-rejection and rejection areas.
 - Critical values are expressed in the same measurement units as the test statistic (z or t).
 - A two-tailed test will have *two* critical values close to the two tails of the curve. Each tail contains $\frac{\alpha}{2}\%$ of the sample distribution means farthest from the hypothesised mean. The critical values or z -scores will correspond to an area $(0.5 - \frac{\alpha}{2})$ from the normal. The critical value in the left tail will take on a negative sign and in the right tail a positive sign.
 - A one-tailed test will have one critical value placed close to one side of the curve. This one tail contains $\alpha\%$ of the sample distribution means farthest from the hypothesis mean. The z -score will correspond to an area of $(0.5 - \alpha)$ from the normal.

Hidden page

Normal z-distribution (if your parameter is a proportion):

$$z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

t-distribution:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

11.1.6 Make the decision

- The **decision rule** is a statement that indicates the action to be taken, that is, do not reject H_0 or reject H_0 .
- Compare the **test statistic** with the **critical value(s)** to see if it falls within the limits of the rejection area or outside the limits.
- If the test statistic falls within the rejection region, the sample evidence does not support the H_0 that the parameter was the specified value and we say reject H_0 .
- If the test statistic falls in the non-rejection region, the sample evidence does support the H_0 that the parameter was the specified value and we say do not reject H_0 .
- We never accept H_0 . Sample evidence can never prove the null hypothesis to be true. When we do not reject the null hypothesis, we are saying that the evidence indicates that the null hypothesis could be true and that there is no statistical evidence to reject it. As long as conclusions are based on sample data, there is a chance that an error could be made.

11.1.7 Interpret the decision

The conclusion should be stated in the context of the original claim. The level of significance should be included and you would say there is enough evidence to support the claim or there is not enough evidence to support the claim.

Example 11.2

A machine is set to fire 30 g of dried fruit into a box of cereal moving along the production line. A sample of 36 boxes revealed that the average mass of fruit inserted was 30.3 g with a standard deviation of 0.5 g. Is the increase in the amount of fruit inserted significant at the 0.01 level of significance?

1. $H_0 : \mu = 30$
2. $H_A : \mu > 30$ (indication of 'more than')
3. $\alpha = 0.01$

The alternative hypothesis uses $>$, implying a right-tailed test.

The central limit theorem applies therefore we use z-distribution.

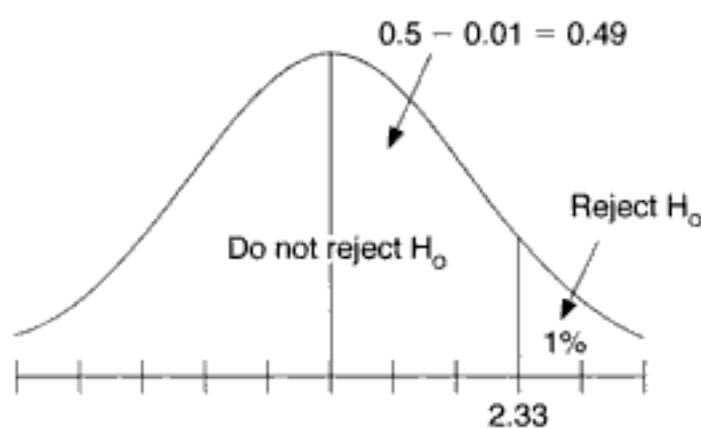
The critical value is 2.33 and the rejection area is: reject H_0 if the z-test > 2.33

4. Test: $z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

$$= \frac{30.3 - 30}{\frac{0.5}{\sqrt{36}}}$$

$$= 3.6$$

5.



6. Since $3.6 > 2.33$ we reject H_0 at the 0.01 level of significance.

7. There is enough evidence to suggest that at a 1% level of significance there is a significant increase in the amount of fruit inserted in a box of cereal.

Activity 11.2

A sample of 100 healthy adult males has a systolic blood pressure of 125 mmHg with a standard deviation of 15. Test at a 2% level of significance whether the mean systolic blood pressure is different from the generally accepted level of 130 mmHg.

Example 11.3

A process takes an average time of 35 minutes. It is thought that a certain modification would reduce this time, and after being modified, the process is repeated 13 times, giving an average time of 33.3 minutes with a standard deviation of 2.4 minutes. Is there any significant reduction in the time at a level of significance of 0.05?

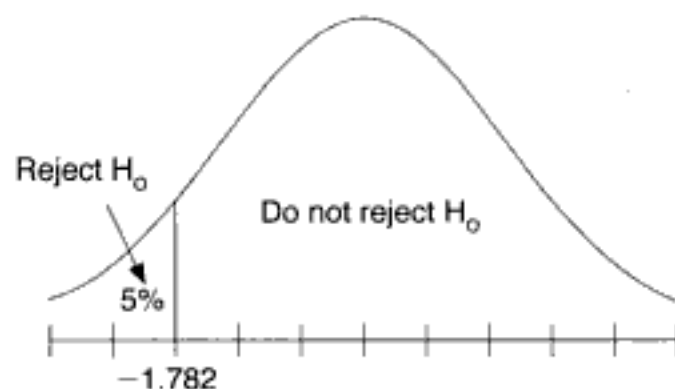
1. $H_0: \mu = 35$

$H_A: \mu < 35$ (reduction in time is an indication of less than)

2. $\alpha = 0.05$

The alternative hypothesis uses $<$, so the test is a one-tail test to the left.

Use the t -distribution because σ is unknown and the sample size is small.



3. To look up the critical t -value, you need to know the direction of the test and the α -value. Find the α in the one-tail test row of the t -table if the test is one-tailed,

and in the two-tail test row if two-tailed. Move down in the chosen column to the required number of *df*. (Remember *df* is $(n - 1)$). The critical *t* is where the *df*-value corresponds with the α .

This example is a one-tail test at a 5% level of significance with 12 degrees of freedom. In the one-tail test row, find the column with heading 0.05. Go down that column to 12 degrees of freedom. The *t*-value that corresponds with this position is 1.782.

Reject H_0 if the *t*-test < -1.782

$$\begin{aligned} 4. \text{ Test: } t &= \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \\ &= \frac{33.3 - 35}{\frac{2.4}{\sqrt{(13 - 1)}}} = -2.45 \end{aligned}$$

- Since $-2.45 < -1.782$ we reject H_0 at the 0.05 level of significance.
- The sample evidence does suggest that there is a significant reduction of the process time.

Activity 11.3

From past records we know that the average unbroken sleep periods of patients with a certain kind of insomnia is 2.8 hours. A new drug is tested on a sample of 25 patients and this yields an average of three hours unbroken sleep with a standard deviation of 0.8 hours. Is there a significant improvement on the unbroken number of hours' sleep? Test at $\alpha = 2.5\%$.

Example 11.4

Directors of a company claim that 90% of the workforce supports a new shift pattern that they have suggested. A random survey of 100 people in the workforce finds 85 in favour of the new scheme. Test at a 5% level if there is a significant difference between the survey results and the directors' claim.

- $H_0: \pi = 0.9$
 $H_A: \pi \neq 0.9$ (no indication of 'more than' or 'less than')
- $\alpha = 0.05$
- The central limit theorem applies therefore we use the normal *z*-distribution. The alternative hypothesis uses \neq , indicating a two-tail test.
- Reject H_0 if the *z*-test > 1.96 or if *z*-test < -1.96

$$\begin{aligned} 5. \text{ Test: } z &= \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} \\ &= \frac{0.85 - 0.9}{\sqrt{\frac{0.9(1 - 0.9)}{100}}} \\ &= -1.67 \end{aligned}$$

Hidden page

1. State the null hypothesis and the alternative hypothesis:

$$H_0: \mu_1 = \mu_2$$

$$H_A: \mu_1 < \mu_2 \text{ (indication of 'less than')}$$

2. Select the level of significance:

$$\alpha = 0.05$$

3. Formulate the decision rule:

The alternative hypothesis uses $<$, so the test is a one-tail test to the left. The central limit theorem applies and we use the z -distribution.

Reject H_0 if the z -test < -1.64

4. Determine the value of the test statistic:

$$\begin{aligned} z &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\ &= \frac{4 - 4.4}{\sqrt{\frac{1.2^2}{40} + \frac{1.5^2}{40}}} \\ &= -1.32 \end{aligned}$$

5. Since $-1.32 > -1.64$ we do not reject H_0 at the 0.05 level of significance.
6. The sample evidence suggests that there is no significant evidence to conclude that the shorter work week does reduce absenteeism.

Activity 11.5

A report on personal savings of 240 citizens of the Gauteng region showed that the average annual savings was R9 300 with a standard deviation of R3 600. The data for a sample of 150 citizens in the Western Cape region showed annual savings of R8 400 with a standard deviation of R2 100. Test at a 5% significance level whether there is a significant difference in the annual savings between the two regions.

Example 11.6

In order to compare if the performance of two training methods are the same, samples of individuals using each of the methods were checked. For the six individuals from method one, the mean efficiency score was 35 with a standard deviation of six. For the eight individuals in method two, the mean efficiency score was 27 with a standard deviation of seven. Set $\alpha = 0.01$.

1. $H_0: \mu_1 = \mu_2$

$$H_A: \mu_1 \neq \mu_2 \text{ (indication of two-tailed)}$$

2. $\alpha = 0.01$

3. The alternative hypothesis sign is \neq , so the test is a two-tailed test.

Both samples are small therefore we use the t -distribution.

If two samples are used, the number of degrees of freedom will be:

$$n_1 + n_2 - 2$$

Reject H_0 if the t -test > 2.681 or if t -test < -2.681

$$\begin{aligned}
 4. \text{ Test: } t &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \\
 &= \frac{35 - 27}{\sqrt{\frac{(6 - 1)6^2 + (8 - 1)7^2}{6 + 8 - 2} \left(\frac{1}{6} + \frac{1}{8} \right)}} \\
 &= 2.24
 \end{aligned}$$

- Since 2.24 falls in the acceptance area, we do not reject H_0 at the 0.05 level of significance.
- The sample evidence suggests that there is no significant difference in the performance of individuals using the two training methods.

Activity 11.6

The manufacturer of two styles of shoes (A and B) wishes to test the hypothesis that the average retail price of style A is less than the average price of style B. A random sample of 12 retailers who stock style A yielded an average price of R146 with a standard deviation of R12. A random sample of 10 retailers who stock style B yielded a mean price of R160 with a standard deviation of R15. Assume that the two samples come from two normally distributed populations and test the hypothesis at a 5% level of significance.

Example 11.7

Workers in two different mine groups were asked what they considered to be the most important labour - management problem. In group A, 200 out of a random sample of 400 workers felt that a fair adjustment of grievances was the most important problem. In group B, 60 out of a random sample of 100 workers felt that this was the most important problem. Would you conclude that these two groups differed with respect to the proportion of workers who believed that a fair adjustment of grievances was the most important problem? Set $\alpha = 0.1$.

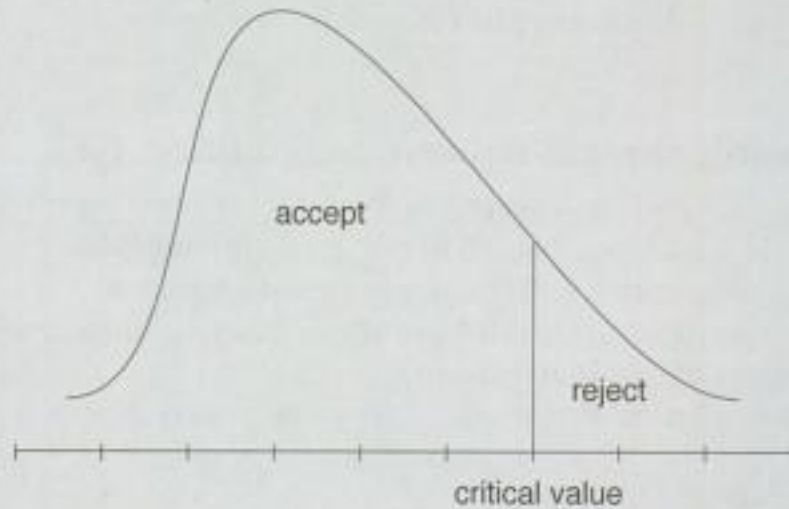
- $H_0: \pi_A = \pi_B$
 $H_A: \pi_A \neq \pi_B$
- $\alpha = 0.10$
- Reject H_0 if the z-test > 1.64 or if the z-test is < -1.64
- $$\begin{aligned}
 z\text{-test} &= \frac{p_1 - p_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \text{ where } \hat{p} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} \text{ and } \hat{q} = 1 - \hat{p} \\
 &= \frac{0.5 - 0.6}{\sqrt{0.52 \times 0.48\left(\frac{1}{400} + \frac{1}{100}\right)}} \\
 &= -1.79
 \end{aligned}$$
- Reject H_0
- There is significant evidence to conclude that the two groups differ in their beliefs.

Hidden page

The top row of the χ^2 -table shows the significance level and the first column contains the number of degrees of freedom. Because the χ^2 -distribution is positively skewed, the critical value will always be positive and in the right-hand tail of the curve.

The acceptance region for H_0 goes from the left tail of the curve to the χ^2 -critical value. To the right lies the rejection area.

You will reject H_0 if the χ^2 -test $>$ χ^2 -table value.



4. Calculate the value of the chi-square test by substituting cell by cell the values from the f_o and f_e table into the formula:

- Construct a table with columns showing the f_o , f_e and χ^2 value for each entry.
- The observed frequencies (f_o) are obtained from the sample data given in the contingency table.
- In order to perform the chi-square test, expected frequencies (f_e) are needed. The (f_e) for any given cell in the contingency table is the product of the total of the frequencies observed in that row and the total of the frequencies observed in that column, divided by the overall size of the sample.

$$f_e = \frac{\text{row total} \times \text{column total}}{\text{overall total}}$$

- The f_o column total should be the same as the f_e column total.
- There is a rule that no f_e should be less than five. When this happens, combine adjacent classes.
- Calculate the chi-square test value using the formula:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

5. Make the decision: in order to determine how different the f_o can be from the f_e and still support the H_0 , the χ^2 -test value is compared with a critical χ^2 -value from the chi-square table.

If the χ^2 -test value exceeds the χ^2 -critical value from step 3, then it falls into the rejection region and H_0 is rejected. If the χ^2 -test value is smaller than the χ^2 -critical value, then H_0 cannot be rejected.

6. Interpret your decision.

Example 11.8

A random sample of adults was selected from each of four ethnic groups in Cape Town. They were asked to specify their primary source of news. The results were as follows:

	Ethnic Group				Total
	A	B	C	D	
TV	30	20	25	20	95
Radio	25	25	20	20	90
Newspaper	10	10	5	30	55
Total	65	55	50	70	240

Is there a relationship between ethnic groups and the source of news at a 2% level of significance?

- H_0 : there is no relationship between ethnic group and source of news
 H_A : there is a relationship between ethnic group and source of news
- $\alpha = 0.025$
- Reject H_0 if the χ^2 -test value > 14.449 (use $df = (k - 1)(r - 1) = (3 - 1)(4 - 1) = 6$)
- Test:

f_o	f_e	$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$
30	25.73	0.71
25	24.38	0.02
10	14.90	1.61
20	21.77	0.14
25	20.62	0.93
10	12.60	0.54
25	19.79	1.37
20	18.75	0.08
5	11.46	3.64
20	27.71	2.15
20	26.25	1.49
30	16.04	12.15
240	240	24.83

- Decision: because the test statistic falls in the rejection region, the H_0 should be rejected at a 0.025 significance level.

6. Conclusion: there is evidence to suggest that there is a relationship between ethnic group and source of news.

Activity 11.8

A manufacturer of women's clothing is interested to know if age is a factor in whether women would buy a particular garment, depending on its quality. A researcher sampled three age groups and each woman was asked to rate the garment as excellent, average or poor. Test the hypothesis, at a 5% level of significance, that rating is not related to age group.

Rating	Age group		
	15 – 20	21 – 30	31 – 60
Excellent	40	47	46
Average	51	74	57
Poor	29	19	37

11.3.2 Goodness-of-fit tests

The χ^2 -goodness-of-fit test for uniform distributions is used to determine whether a set of sample data differs significantly from what is expected.

Steps

1. State the null and alternative hypotheses:
 H_0 : The population under investigation fits some specified or expected distribution.
 H_A : The population does not fit the specified distribution
2. Select the level of significance: α is the criterion used to formulate the rejection area for H_0 .
3. Define the rejection region: to find the critical values from the chi-distribution table, you need the level of significance (α) and the degrees of freedom:
 $df = k - 1$ where k is the number of possible outcomes in the investigation.
 Reject H_0 if the χ^2 test statistic $>$ χ^2 -critical value.

4. Calculate the χ^2 -test statistic: $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$

Where:

f_o is the observed frequency from the sample data, and

f_e is the expected frequency that is calculated to conform to the null hypothesis that is being tested.

If the calculated test statistic is zero, it means that the observed frequencies and expected frequencies are identical, or exactly what we had expected.

5. Make the decision.
6. Interpret the decision.

Example 11.9

A manufacturer of soap wishes to know if consumers have a preference for bath soap fragrances. To answer their question, a random sample of 200 adult shoppers is offered a free bar of soap. The recipients may choose from among four flavours. The choices are as follows:

Rose	Lavender	Sandalwood	Lemon
66	53	45	36

1. H_0 : there is no preference, i.e. all flavours are equal
 H_A : there is a preference in respect of flavour, i.e. flavours are not equal.
2. Level of significance: $\alpha = 0.01$
3. Critical value: $\chi^2 = 11.345$ using $df = k - 1 = 4 - 1 = 3$
 ($k =$ number of groups or intervals)
 Reject H_0 if the χ^2 -test > 11.345
4. Test:

Flavour	f_o	f_e	$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$
Rose	66	50	5.12
Lavender	53	50	0.18
Sandalwood	45	50	0.50
Lemon	36	50	3.92
Total	200	200	9.72

- f_e : if we expect no preference, the expected frequency for all flavours should be equal ($200 \div 4 = 50$).
 - The f_o column total should be the same as the f_e column total.
 - There is a rule that no f_e should be less than five. When this happens combine adjacent classes.
5. Decision: the test statistic < 11.345 and it falls in the acceptance region. There is no evidence to reject the H_0 at a 0.01 significance level.
 6. Conclusion: there is no evidence to suggest that there is a preference with respect to fragrance.

Example 11.10

The respective car manufacturer's shares of the national market are as follows:

Manufacturer	% of shares
Volkswagen	37
Toyota	30
Delta	15
BMW	10
Mercedes	8

A random sample of 2 000 car owners in Pretoria revealed the following ownership pattern: Volkswagen 758, Toyota 680, Delta 300, BMW 162 and Mercedes 100. Does the ownership pattern in Pretoria differ significantly from the national pattern?

- H_0 : the ownership pattern in Pretoria is the same than the national pattern.
 H_A : the ownership pattern in Pretoria differs from the national pattern.
- $\alpha = 0.05$
- Reject H_0 if the χ^2 -test > 9.488 ($df = k - 1 = 5 - 1 = 4$)
- Test:

Manufacturer	National pattern (f_e)	Pretoria pattern (f_o)	$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$
Volkswagen	37% = 740	758	0.44
Toyota	30% = 600	680	10.67
Delta	15% = 300	300	0
BMW	10% = 200	162	7.22
Mercedes	8% = 160	100	22.5
	2 000	2 000	40.83

- Decision: because the test statistic falls in the rejection region, the H_0 should be rejected at a 0.05 significance level.
- Conclusion: there is evidence to suggest that the pattern in Pretoria differs from the national pattern.

Activity 11.9

A delivery of assorted nuts is labeled as having 45% walnuts, 20% hazelnuts, 20% almonds and 15% brazil nuts. By randomly picking several scoops of nuts from the bag delivered, the following count was obtained:

Walnuts	Hazelnuts	Almonds	Brazil nuts
92	69	32	42

Could these findings be a basis for an accusation of mislabeling at a 2.5% level of significance?

TEST YOURSELF 11

1. Frequent checks were made of the spending patterns of tourists returning from countries in Asia. Results indicated that travellers spent an average of R1 010 per day. In order to determine whether there has been a change in the average amount spent, a sample of 70 travellers was selected and the mean was determined as R1 090 per day with a standard deviation of R300. Is there evidence of a significant increase in the mean amount spent per day at the 0.01 level of significance?
2. The desired percentage of silicon dioxide in a certain type of cement is 5.0. A random sample of 36 specimens gave a sample average percentage of 5.21 and a sample standard deviation of 0.38. Use a significant level of 0.01 and test whether the sample result indicates a change in the average percentage.
3. A nutritionist claims that the mean tuna consumption by a person is 1.55 kg per year. A sample of 60 people shows that the mean tuna consumption by a person is 1.45 kg per year with a standard deviation of 0.51 kg. At $\alpha = 0.02$, can you reject the nutritionist's claim?
4. A machine is set to fire 30 g of dried fruit into a box of cereal moving along the production line. A sample of 36 boxes revealed that the average mass of fruit inserted was 30.3 g with a standard deviation of 0.5 g. Is the increase in the amount of fruit inserted significant at the 0.05 level of significance?
5. A company that makes cola drinks states that the mean caffeine content per bottle of cola is 40 mg. The quality controller is convinced that it is lower. A sample of 30 bottles of cola has a mean caffeine content of 39.2 mg with a standard deviation of 7.5 mg. At $\alpha = 0.01$, can the quality controller reject the claim?
6. Hyperactive children are often disruptive in the typical classroom setting because they find it difficult to remain seated for extended periods of time. The typical number of 'out-of-seat' behaviours was 12.40 per hour. Treatment was applied to a group of 25 hyperactive children and after treatment the 'out-of-seat' behaviours reduced to 11.60 per hour with a standard deviation of 3.5. Using $\alpha = 0.01$, can we conclude that this decline is significant?
7. Medical research has shown that repeated wrist extension beyond 20° increases the risk of wrist and hand injuries. In each of 24 randomly selected students in the information technology field, the wrist extension was recorded while using a mouse with a proposed new design. The sample mean was found to be 24° with a standard deviation of 5° . Test the hypothesis that the mean wrist extension for people using the new mouse design is greater than 20° .

8. You are involved in an environmental awareness program and want to test the claim that the mean waste generated by adults is more than 1.8 kg per day. In a random sample of 15 adults, you find that the mean waste generated per person per day is 1.9 kg with a standard deviation of 0.54 kg. At a 5% level of significance, is the claim justified?
9. A random sample of 16 unflavoured ice-cream tubs were selected at random and subjected to chocolate flavouring. The sample mean time required to flavour the ice-cream was 13 minutes with a standard deviation of two minutes. Perform a hypothesis test at the 1% level of significance to test that the population mean time required to flavour ice-cream is greater than 10 minutes.
10. A chicken producer claims that the average mass of a particular group of chickens is 1 kg. Before agreeing to purchase, a customer selected a sample of 25 chickens, which yielded a sample mean of 1.12 kg and standard deviation of 0.1 kg. If the masses can be considered to be normally distributed, should the claim be rejected at the 1% level of significance?
11. A personnel manager claims that 60% of all single women hired for secretarial jobs leave to get married within two years. An analysis shows that of a random sample of 120 single women, 64 left to get married. Is this evidence consistent with the company's claim, at a 1% level of significance?
12. A plant is producing large numbers of water testing equipment of which, on average, 2% are defective. In a random sample of 1 000, 3% are found to be defective. Does this indicate a significant deterioration in the process? Test at a level of significance of 0.02.
13. A company manufacturing salad dressings claimed that 85% of households eat salad at least once a week. A nutritionist suspects that the percentage is higher than this. She sampled 200 households and found that 170 of them eat salad at least once a week. Conduct a test to address the nutritionist's suspicions. Use $\alpha = 0.10$.
14. Club 60 claim that senior citizens participating in some sort of exercise have a blood pressure lower than the average of 160 mmHg. To test this claim, 20 active senior citizens were selected at random and their blood pressure was found to average 151 with a standard deviation of 12. Is Club 60's claim valid at a 10% level of significance?
15. A manufacturer claims that his market share is 60%. However a random sample of 500 customers reveals that only 275 are users of his product. Test the claim at the 2% level of significance.
16. The sales manager wants to determine if the average size of orders received by the company's eastern branch differs significantly from the average size of orders received from the western branch at a 2% level of significance. A random sample of 90 orders from the eastern branch had a mean value of R131.60 with a standard deviation of R25.80. A random sample of 55 orders received by the western branch had a mean value of R115.70 with a standard deviation of R32.23.
17. A large bank is affiliated with both the Mastercard and Visa credit cards. For a sample of 100 Mastercard holders, it is observed that the average month-end account balance is R680 with a standard deviation of R300. For a random sample of 100 Visa cardholders, the average month-end account balance is R550 with a standard deviation of R265. Is the average for the Visa card holders significantly lower than the Mastercard average? Test at $\alpha = 0.10$.

18. A consumer testing service compared gas ovens to electric ovens by baking one type of bread in five ovens of each type. The gas ovens had an average baking time of 0.9 hours with a standard deviation of 0.09 hours and the electric ovens had an average baking time of 0.7 hours with a standard deviation of 0.16 hours. Test the hypothesis that the baking times are the same in both kinds of ovens at the 5% level of significance. Assume the baking times are normally distributed.
19. In order to conduct a consumer behaviour survey, a sample of 500 residents was selected in a metropolitan area. One of the questions asked was "Do you enjoy shopping for clothing?" Of 240 males, 136 answered yes. Of 260 females, 224 answered yes. Determine whether there is evidence that the proportion of females who enjoy shopping for clothing is higher than the proportion of males, using a 5% level of significance.
20. In an experiment to compare the fracture toughness of high-purity steel with commercial-purity steel of the same type, 32 specimens were selected from each type. The sample mean and standard deviation toughness for the high-purity steel specimens were 65.6 and 1.4 respectively. The sample mean and standard deviation toughness for the commercial-purity steel specimens were 59.2 and 1.1 respectively. Test at a 5% level of significance whether a significant difference exists between the two types.
21. A supermarket chain is interested in determining whether a difference exists between the mean shelf life (in days) of two different brands of bread. Random samples of 50 freshly baked loaves of each brand were tested with the results shown below:

	Brand A	Brand B
Sample mean	4.1	5.2
Sample standard deviation	1.2	1.4

Is there sufficient evidence to conclude that brand B has a longer shelf life than brand A at a 2% level of significance?

22. In a public opinion survey, 60 out of a sample of 100 high-income voters and 40 out of a sample of 75 low-income voters supported a decrease in VAT. Can we conclude at a 5% level of significance that the proportion of voters favouring a decrease differs between high- and low-income voters?
23. In an Aids-awareness program, it was found that 110 males in a random sample of 310 males were aware of Aids. In another similar program, it was found that 87 women in a random sample of 290 women were aware of Aids. Test at the 2% level of significance whether the first campaign was more successful.

24. Tests have been carried out on the effects of three fertilisers on sugar cane growth. Each fertiliser was tried on several different plots of land. Each value is a number of plots of land.

	Fertiliser		
	A	B	C
Strong growth	94	124	44
Weak growth	50	96	38

Test for an association between the choice of fertiliser and plant growth at a 1% level.

25. A car manufacturer is interested in predicting purchase patterns for a new small capacity car they are producing. The car comes in four colours and the manufacturer wants to relate colour preference to the gender of the purchaser. Use the following sample data and do the hypotheses tests at a 10% level of significance:

	White	Green	Red	Silver
Male	260	240	175	420
Female	130	200	240	340

26. Two different manufacturers supply parts for a production process. Each part is tested for six possible defects. The following table shows the number of each type of defect by each supplier:

Supplier	Defect					
	1	2	3	4	5	6
A	35	10	10	2	5	10
B	45	20	0	10	15	20

Would you conclude that the defect is independent of the supplier, using a 2.5% level of significance?

27. A sales manager has become interested in the number of sales calls made by each of the employees. He reasoned that if all the employees are working equally hard, they should make the same number of calls during a set period of time. In order to investigate this hypothesis, the manager used a sample of five employees and recorded the number of calls they made during a set time period:

Employee	A	B	C	D	E
No. of calls	31	62	59	40	58

At a 1% level of significance, is the manager's idea supported?

Hidden page

32. Supermarket chains often carry products with their own brand labels and usually price them lower than the other brands. A supermarket conducted a taste test to determine whether there was a difference in taste among the four brands of ice cream it carries: own brand (A) and three other brands B, C and D. A sample of 200 people participated and they indicated their preference as shown in the table below:

A	B	C	D
39	57	55	49

Test at a 5% level of significance if there is a difference in preference for the four brands.

PART
2

Calculation
Skills

UNIT 12

Elementary calculations

In this unit you revise your calculation skills and how to utilise your calculator.

After completion of this unit you will be able to:

- classify the numbers you are dealing with
- understand common mathematical notation
- apply rounding rules to the results of calculations
- deal with additions, subtractions, multiplications and divisions
- deal with signed numbers
- understand and use exponents, square roots, logarithms, factorials and summation
- deal with fractions, decimals and the metric system.

The purpose of this course is to provide you with the numeracy skills to understand the basic principles of business calculations and make sound decisions based on them.

These skills will benefit you in other subjects, in a business career, and even in the everyday business of living.

12.1 The electronic calculator

1. Power switch: all calculators need to be switched ON; the power supplied by the batteries or electricity allows the user to enter, display, calculate and store values. When a calculator is switched on, the display screen should light up and display a set of numbers, usually a zero or a zero and a number of zeros after the decimal point. A feature of most modern calculators is that the display becomes blank after several minutes of non-use. This is due to an automatic 'power-off' function, which is designed to prolong the life of the battery by turning off the display. Once this occurs, power can be restored either by switching the power of the unit OFF and then ON again, or by pressing the AC key.
2. The face of a calculator normally consists of two parts, namely:
 - The display screen, where numbers and/or letters are digitally displayed for you to read. The display window shows calculation values and results. Some calculators have a two-line display showing the calculation formula on the first line and its answer on the second line.
 - The keypad, where there are a number of keys (buttons) through which the calculations are performed. Depending on the type of calculator used, every

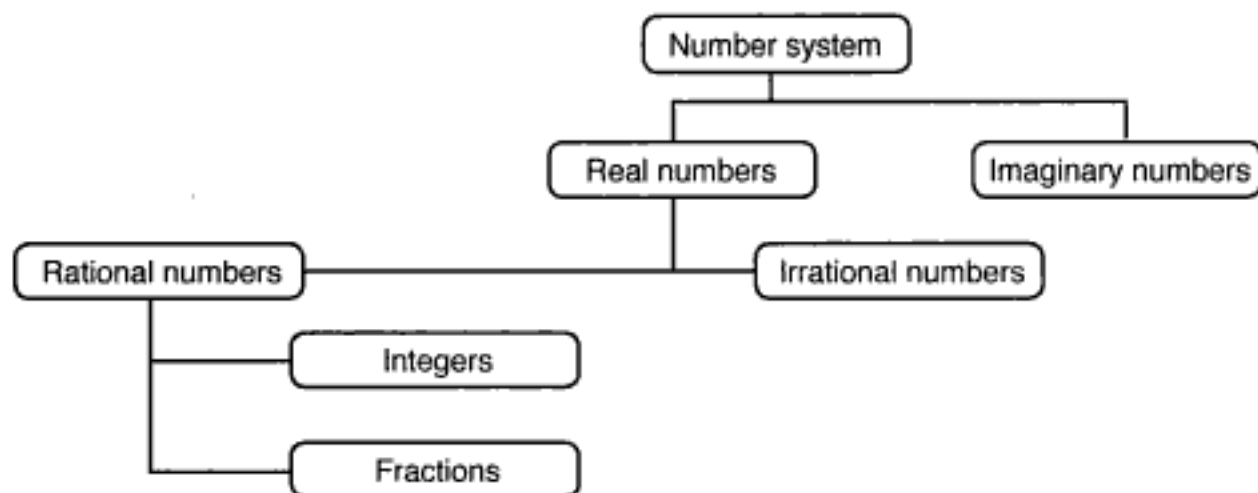
key has an inscription on it, indicating the function it performs. There may also be other inscriptions on the keypad itself, above or below the different buttons. These inscriptions are usually coloured differently, indicating that more than one function is assigned to that key (the same button can perform another function).

- To indicate to the calculator that it is the other function that you wish to invoke, the SHIFT, 2nd, or INV key has to be pressed prior to pressing the key itself. This method of multiple functions assigned to individual keys helps to keep a calculator small, compact and versatile.
3. The clear or cancel key:
 - Always make sure that you 'clear' the display screen before starting a new calculation by touching the 'cancel' (C , CLR or AC) key on the calculator.
 - If you enter a series of values for a calculation and have typed an incorrect value, you correct it without destroying the previous intermediate calculations, by pressing the C key, or on some calculators an arrow: ◀ or ▶. It clears the last entry and waits for you to continue.
 - The AC key clears any pending calculations and resets the display to '0'. Any values in the other memory locations are unaffected by it.
 - Data held in the independent memory (M), as well as MODE specifications, are held in memory even when power is turned OFF . Pressing the SHIFT, 2nd , or INV key followed by the C or AC key clears the storage memory locations.
 4. The MODE key: the keys on a calculator perform different operations depending on the mode entered. Pressing the MODE key and depressing a number from 1 to ..., depending on the number of modes available on the calculator, sets the specific mode required. The different calculation modes are entirely independent, and cannot be used in combination with each other. Some MODES available on calculators are: normal calculations, standard deviation calculations, regression calculations and interest calculations.
 5. The number of digits that can be entered into a calculator depends on the size of the display, normally 10 digits. When the resulting answer exceeds the display limit, the value is displayed in scientific notation. The display reads as follows in these cases:
 - 1.4^{05} is 140 000. The 05 to the right of the value means that the decimal point must move 5 places to the right.
 - 1.4^{-04} is 0.00014. The decimal point is moved 4 places to the left.
 6. The different keys to use for specific calculations will be mentioned when the operation is dealt with in the module. It is also recommended that you keep your calculator's owner's manual or user's guide accessible, because different calculators use different methods, which your trainer may not be familiar with.

12.2 The number system

The number system we use today is known as the Hindu-Arabic number system.

12.2.1 Classification of numbers

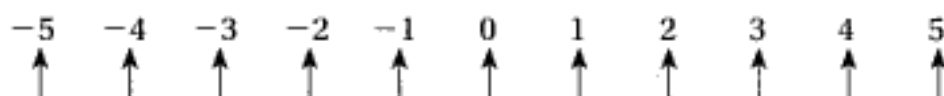


12.2.2 Real numbers and imaginary numbers

Numerical values can be classified as **real numbers** and **imaginary numbers**.

Imaginary numbers are the square roots of negative numbers, such as $\sqrt{-2}$, and it is difficult to attribute any real or practical significance to them. They form part of the complex number system, which is beyond the scope of this text.

The number system that enables us to assign a number to every possible point on a number line is called **real numbers**. A number line is a scaled straight line with a zero point, on which we can indicate the position of numbers and their interrelations. Negative numbers are indicated to the left of the 0 and positive numbers to the right of the 0. The number line can extend to infinity with fractions or decimals between the whole numbers.



Real numbers consist of **rational numbers** and **irrational numbers**.

Activity 12.1

Fill in the missing numbers in each of the following rows:

	66	67				71		73
3		9						27
	30		40				60	
100		120						
		36	48					

12.2.3 Rational numbers and irrational numbers

A **rational number** can be expressed as the ratio or fraction $\frac{a}{b}$ in which both numerator and denominator are whole numbers or a decimal that repeats or terminates, and $b \neq 0$.

In the fraction $\frac{a}{b}$, a is called the numerator and b is called the denominator.

Division by 0 is undefined.

The three groups of rational numbers are:

1. **Integers or whole numbers.** For example: 6 is a rational number because it can be written as $\frac{6}{1}$ (both numerator and denominator are whole numbers).
2. Finite or **terminating fractions.** For example: $\frac{2}{8}$ is a rational number because the decimal expression 0.25 is terminating or finite.
3. Recurring or **repeated fractions.** For example: $\frac{2}{3}$ is a rational number because the decimal expression 0.666... has a pattern that repeats or can carry on forever. This recurring decimal can also be written as $0.\dot{6}$.

Irrational numbers are real numbers which cannot be expressed as the ratio of two integers. The decimal value cannot be expressed with either a finite number of decimal places or a repeating pattern. The never-ending string of digits will not form a pattern that continues to repeat itself. Any irrational number falls between two rational numbers.

Example 12.1

Recurring decimals

$\sqrt{5} = 2.2360679775\dots$ This is an irrational number because it cannot be written as $\frac{a}{b}$ and this never-ending string of digits will not form a pattern that continues to repeat itself.

Activity 12.2

Classify the following numbers as rational and irrational.

1. $\sqrt{2}$
2. $\frac{5}{7}$
3. $\frac{3}{13}$
4. $\frac{13}{19}$

Note: The distinction between rational and irrational numbers is of very little significance as far as practical applications are concerned. This is due to the fact that any irrational number can be approximated (rounded) to any desired degree of accuracy by means of a rational number.

12.2.4 Whole numbers and fractions

An **integer** is a positive whole number (1, 2, 3, ..., also known as a natural or counting number), a negative whole number (-1, -2, -3, ...) and the value zero (0). The number zero is neither negative nor positive, and in that sense is unique. The zero point is called the origin of the number system. **Integers** are rational numbers, because an integer n can be considered as the ratio $n/1$.

Fractions, such as $\frac{1}{2}$, $\frac{3}{7}$ or $\frac{1}{3}$, will fall between the integers on the number line.

Fractions can be classified as *proper fractions*, such as $\frac{1}{2}$, $\frac{3}{7}$ or $\frac{3}{4}$, or *improper fractions* such as $\frac{4}{2}$, $\frac{13}{7}$ or $\frac{13}{4}$ where the numerator is always bigger than the denominator.

A *mixed number*, such as $1\frac{1}{2}$, $\frac{13}{7}$ or $5\frac{3}{4}$, is a whole number with a fraction. Any mixed number can be turned into an improper fraction.

To change a fraction into a decimal, divide the numerator by the denominator.

Note: A natural number is a counting number beginning with 1.

Activity 12.3

In the table below tick (✓) in the column(s) that correctly describe the number as real, irrational, rational, integer, whole or natural. Use a calculator to help you.

	Real number	Irrational number	Rational number	Natural number	Whole number
$\sqrt{5}$					
54 876					
-16					
$\frac{5}{8}$					
1.25					
$(\frac{1}{2})^3$					
$3\frac{1}{7}$					

12.3 Common notation

Mathematical 'shorthand', or symbols in analysing and presenting results, is often used rather than descriptive text.

Arithmetic symbols:

+	add	-	subtract
×	multiply	÷	or / divide
<	less than	≤	less than or equal to
>	more than	≥	more than or equal to
=	equal to	≠	not equal to
±	plus/minus	Σ	sum of
≈	rounded as	n!	factorial

12.4 Basic operations

12.4.1 Hierarchy in calculations

The order of operations is a set of rules that mathematicians have agreed to follow to avoid mass confusion when simplifying mathematical expressions or equations. When more than one calculation has to be made to find the solution of a mathematical expression, you must follow a specific order of operations. The following are the priority order of operations:

1. Exponents and roots are all treated as functions. Turn the functions into numbers first.
2. Perform any calculations inside brackets ().
3. Do all powers and roots.
4. Complete all multiplication and division, working from left to right.
5. Perform additions and subtractions, working from left to right.

Note: To change the order of priority, brackets or the calculator can be used. Note that the multiplication symbol '×' is frequently omitted in some expressions. For example: $6 \times (5 - 2)$ will normally be shown as $6(5 - 2)$.

Activity 12.4

Do the following calculations:

1. $2 \times 6 + 3 - 4/2 - 5 + 20/5 \times 3 + 50$
2. $(3 + 3 - 5)(15 - 5)10 - \log 99$
3. $4 + 5 - 7 + 8 \times 5 - 12 \times 2/8 + 6 - 3 + 20 \div 2$
4. $9 \times 9 - 30 \div 3 + 5 - 6 + 7 - 2 + 9 \times 9$
5. $[10 \times 4 - 6 + 7 - 8/2 + 3 \times 3 + (4 + 5 - 6/3) + 1]/2$
6. $[(3 + 4 - 6/2 + 2) + (9/3 + 6 \times 5)/11] \times [(4 + 5 - 6) + (18 - 3 \times 4)]/9$
7. $2 + 3 \times 15 - 10/2 + (5 + 1)/3$
8. $30/15 - 10$
9. $30/(15 - 10)$
10. $(40 \times 2 - 50/2) \times (9/3)/100$

12.4.2 Adding and subtracting (+ and -)

Addition and subtraction are the most common of the fundamental operations and are easy to perform. Adding and subtracting with speed and accuracy can be achieved only through practice. Adding is to add more (or calculate the sum of) and subtracting is to take away (or calculate the difference).

Note: The order of operations (adding or subtracting) does not matter.

Activity 12.5

1. Do the following calculations:
 - a) $28 + 5 + 3 + 6 =$
 - b) $16 - 7 - 2 =$
 - c) $7 + 46 - 15 + 3 =$
2. The distance from Johannesburg to Polokwane is 340 kilometres. The distance from Johannesburg to Pretoria is 59 kilometres. How far is it from Pretoria to Polokwane?
3. The following number of items had to be scrapped at the end of the day due to damage: 21 shirts, 32 pairs of pants, 10 pairs of shoes and 53 jerseys. How many items had to be scrapped?
4. Harry worked 15 hours of overtime in January, 12 hours of overtime in February and four hours of overtime in March. What is the total number of hours Harry has worked overtime?
5. There are 11 people in a taxi. At the next stop five people get off and nine get on. How many people are now in the taxi?
6. You have 57 items in stock at the beginning of a shift. During the shift you sell 33 items and receive a new delivery of 25 items. How many items will you have in stock at the end of the shift?
7. If you want to buy a new suit for R599.99 and you have only R315 available in your account, how much must you pay into your account to be able to buy the suit?
8. If you left at 07h00 in the morning and arrived at 12h45 that afternoon, how long did the journey take?

12.4.3 Multiplying and dividing (\times and \div)

The arithmetic operation to determine the product of numbers is called multiplication (\times).

Division (\div) is the arithmetic operation that finds how many times one number goes into another.

Note: The order in which you multiply and divide does not matter, but, if the calculation includes addition and subtraction, you must first calculate values inside brackets, or multiply and divide, before you add and subtract.

Note: An alternative to the multiplication sign (\times) is the multiplication point (\cdot), a point that is set above the line and not to be confused with the decimal point.

An alternative to the division sign (\div) is the right oblique ($/$), as used in writing fractions.

Example 12.2

- You have four boxes with 24 bars of soap in each and you want to know how many bars of soap you have in total:
 $4 \times 24 = 96$.
- You must pack 100 items in boxes containing five items each and you want to know how many boxes you need:
 $100 \div 5 = 20$ boxes.

Activity 12.6

- Do the following calculations:
 - $379 \times (-15) =$
 - $69 \div 13 =$
 - $36 \div 6 \times 5 =$
- If the shop is open from 08h00 in the morning till 21h00 at night, how many hours must each employee work if you have 2 shifts during the day?
- Abel takes 10 minutes to unpack one box of shirts. If there are 25 boxes, how long will it take him to unpack all the boxes? If there are eight shirts per box, how much time does he spend to unpack each shirt?
- Deon's wage is R25 per hour. If he works eight hours per day for five days and only four hours on Saturday, what is his weekly wage before tax?
- If you buy nine pairs of socks at R14.99 per pair, how much change should you get if you pay with a R200 note?
- You need 250 g of flour to bake a loaf of bread. If you have a 5 kg packet of flour, how many loaves of bread can you bake?
- Your truck can take at most four tons at a time. If you want to move 38 tons of clothing to the warehouse, how many trips will you have to make? If each trip takes two hours and 15 minutes, how many hours will you take to move the clothing? How many tons will you take on each trip?
- You want to put a carpet in the staff room. You have to pay R55 per m^2 . The room is 11 m long and 6 m wide. How many square metres of carpet do you need? How much will it cost you?

12.5 Signed numbers

1. When adding numbers of the same sign, find the sum of the numbers and use the sign common to all factors.

$$2 + 3 + 4 = +9$$

$$(-2) + (-3) + (-4) = -9$$

2. When adding numbers of different signs, find the sum of the positive numbers and the sum of the negative numbers, and then subtract the smaller sum from the bigger one and designate the sign of the bigger sum.

$$(-2) + 3 + 4 + (-1) = +7 - 3 = +4$$

$$2 + (-3) + (-4) + 1 = +3 - 7 = -4$$

3. When subtracting a negative number, change the sign of the negative number being subtracted and add the number to the rest.

$$2 + 3 - (-4) = +9$$

4. When multiplying or dividing by the same sign, the answer will always be positive.

$$-2 \times -2 = +4$$

$$2 \times 2 = +4$$

$$-2 \div -4 = +0.5$$

5. When multiplying or dividing unlike signs, the answer is always negative.

$$-2 \times 2 = -4$$

$$3 \times -4 = -12$$

$$-6 \div 2 = -3$$

Note: Use the $(-)$ or \pm key on the calculator to change the sign of a number.

Activity 12.7

Do the following calculations:

1. $(13) + (-2) =$

2. $(-2) - (+3) =$

3. $(-4) + (-7) =$

4. $0 - (-3a) =$

5. $(-3) + (+1) =$

6. $(+5) \times (+1) =$

7. $(-xy)(-1) =$

8. $(-12xy) \div (-4) =$

9. $(12t)(-t) =$

10. $(-5p) \cdot (12p) =$

12.6 Exponents (powers) (x^y)

When a value is multiplied by itself some number of times, a superscript number can be placed at the upper right-hand side of the value.

For example: $2 \times 2 \times 2 \times 2 = 2^4$ (read as two to the power of four). The superscript number (4) is known as the exponent and 2 is the base. If a number is raised to the power of 2, it is known as the square of the number.

Hidden page

Activity 12.9

Calculate the following expressions accurate to the nearest hundredth

1. $\text{Log } 340 =$
2. $1 + 3.3\text{Log } 50 =$
3. $e^3 =$
4. $e^{-12} =$

12.9 Factorial notation (!)

The symbol $n!$ is read as n factorial, and is a shorthand way of identifying the product of all the positive numbers from 1 up to n .

For example: $5! = 5 \times 4 \times 3 \times 2 \times 1$

$$10! = 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$$

Notes:

- The symbol $n!$ has no meaning if n represents anything other than a positive whole number or zero. ($-3!$ is undefined)
- The value of $0!$ is defined to be 1. ($0! = 1$)
- To obtain the factorial value from the calculator, enter the value followed by the $x!$ or $n!$ key.

Activity 12.10

Do the following calculations:

1. $9! =$
2. $\frac{13!}{10!} =$
3. $\frac{20!}{8!12!} =$

12.10 Sigma notation (Σ)

This Greek capital letter sigma (Σ) stands for 'sum the appropriate values'.

Thus we write $1 + 2 + 3 + 4 + \dots + n$ as $\sum_{i=1}^n x_i$

This means the sum of all the x values from 1 through n . This index system must be used whenever only part of the available information is to be used. In statistics, however, we will usually use all the available information and the notation will be adjusted by doing away with the index system.

$\sum_{i=1}^n x_i$ will become Σx if all the data are used.

Activity 12.11

Use the x -values 5 3 2 6 3 to calculate:

1. $\sum x =$
2. $\sum x^2 =$
3. $(\sum x)^2 =$

12.11 Fractions

A fraction is a number that can represent part of a whole and is denoted by: $\frac{a}{b}$ if a and b are integers and $b \neq 0$.

- The numerator (a) at the top tells you how many parts of the whole are actually used.
- The denominator (b) at the bottom tells you how many parts the whole has been divided into. If the numerator is more than the denominator, we have an improper fraction. If the numerator is smaller than the denominator, we have a proper fraction.
- A horizontal line drawn between the numerator and denominator separates them. This horizontal line indicates that the numerator value must be divided by the denominator value to find a single numerical result. A decimal can thus be found by dividing a numerator by a denominator.
- Converting the fraction to a decimal before performing the arithmetic operation can save computational time.

Example 12.3

The fraction $\frac{3}{4}$ indicates 3 is to be divided by 4 as follows:

$$3 \div 4 = 0.75 \text{ or } \frac{3}{4} = 0.75$$

Rules governing fractions

- Any fraction in which the denominator equals the numerator has a value of one.
- Division into 0 is mathematically undefined and the fraction will always equal 0.
- When multiplying or dividing both the numerator and the denominator by the same value, the fraction value does not change.

$$\left\{ \frac{1}{2} \cdot \frac{5}{5} = \frac{1}{2} \right\}$$

- A fraction can be reduced to its lowest terms by dividing both the numerator and denominator by a common factor.

$$\left\{ \frac{5 \div 5}{10 \div 5} = \frac{1}{2} \right\}$$

- To add or subtract fractions with the same denominator, add or subtract the numerators and write the sum over the common denominator.

$$\left\{ \frac{3}{5} + \frac{1}{5} = \frac{4}{5} \right\}$$

- To add or subtract fractions with different denominators, a common denominator is found by multiplying the denominators and restating each fraction with the common denominator.

$$\left\{ \frac{5}{8} + \frac{1}{3} = \frac{5(3) + 1(8)}{24} = \frac{23}{24} \right\}$$

- To multiply fractions, multiply both the numerators and denominators.

$$\left\{ \frac{2}{5} \times \frac{3}{7} = \frac{6}{35} \right\}$$

- To multiply a fraction by a whole number, multiply the numerator by the whole number, maintaining the same denominator.

$$\left\{ 3 \cdot \frac{1}{2} = \frac{3}{2} = 1\frac{1}{2} \right\}$$

- When dividing by a fraction, inverse the fraction in the denominator and then multiply the numerator by the inverted fraction.

$$\left\{ \frac{\frac{5}{12}}{\frac{3}{4}} = \frac{5}{12} \times \frac{4}{3} = \frac{20}{36} \right\}$$

Note: Fraction calculations can be done on the calculator if the fraction key $\frac{b}{c}$ is available.

Activity 12.12

- Convert the following fractions to their decimal equivalents. Round your answer to two decimal places.

$$\frac{1}{3} = \quad \frac{23}{8} = \quad 3\frac{3}{19} = \quad \frac{4}{11} =$$

- Convert the following decimals to fractions:

$$0.11 = \quad 0.135 = \quad 0.1567 = \quad 2.1723 = \quad 0.07 =$$

- Susan earns an hourly wage of R25 per hour. If she works overtime she gets $1\frac{1}{2}$ times her wage for the first five hours and $1\frac{1}{3}$ after the first five hours of overtime.

- What is her hourly wage for the first five hours of overtime?
- What is her hourly wage if she works longer than five hours of overtime?
- How much does she earn before tax if she works 13 hours overtime?

- Give the following answers in fractions:

- $\frac{1}{3} + \frac{3}{4} =$
- $\frac{5}{20} - \frac{1}{5} =$
- $\frac{2}{6} \times \frac{3}{9} =$
- $\frac{5}{24} \div \frac{1}{6} =$

12.12 Decimal numbers

Each digit in a number has a place and a value. The location of a digit within a number is called its place. Place value is the value of the digit based on its location within a number group. Our numerical (and monetary) system is based on the 10 numbers (decimal) metric system and as such, decimal fractions always have a power of 10 in the denominator (e.g. 10, 100, 1000).

- A fraction can be converted to a decimal if the numerator is divided by the denominator.
- Decimals have a decimal value relative to their position in relation to the 'decimal point' in a number.
- Numbers to the left of the decimal point are whole numbers and numbers to the right of the decimal point have values of tenths, hundredths, thousandths etc. of a whole number.

Example 12.4

The following example shows the place name of each digit:

5	4	3	2	3	6	4	.	3	1	0	8	7	5
Millions	Hundred-thousands	Ten-thousands	Thousands	Hundreds	tens	units	Decimal	Tenths	Hundredths	thousandths	Ten-thousandths	Hundred-thousandths	millionths
M	Thousands			Ones			.	Decimals					

There are 5 millions (5 000 000), 4 hundred thousands (400 000), 3 ten thousands (30 000), 2 thousands (2 000), 3 hundreds (300), 6 tens (60), 4 units (4), 3 tenths (0.3), 1 hundredth (0.01), 0 thousandths (0.000), 8 ten-thousandths (0.0008), 7 hundred-thousandths (0.00007), 5 millionths (0.000005).

Activity 12.13

Complete the following:

832 means:	2 units, 3 tens and 8 hundreds.
611 means:	
1 093 549 means:	
3.026 means:	
522.034563 means:	

Hidden page

12.14 Rounding off decimals

It is often desirable to round numbers to make them easier to understand and use. Round numbers after completing arithmetic operations and not before, and ensure that results are acknowledged as such.

- Draw a line after the digit you want to round, for example if you want to round off the number to two decimal places, you draw a line after the second decimal.
- If the digit to the right of the desired rounding digit has a value of five or more, increase the rounding digit by one and replace succeeding digits with zeros if it is a whole number or disregard all numbers to the right if decimals.
- Should the value to the right of the desired rounding digit be less than five, leave the rounding digit as it is and disregard all numbers to the right of the cut-off digit or replace with zeros if it is a whole number.
- If the digit to the right of the desired rounding place is the last digit and exactly five, increase the rounding digit by one if it is an odd number and leave as is if the number to its immediate left is an even number.
- Often, a calculation results in more decimal places than is necessary. There may however not be more decimals to the right of the decimal point than in the values being processed – that means you can't be more accurate than the least accurate value of the given data. To correct such an instance requires a 'rounding off' to the appropriate number of decimal places, for example $1.2 \times 0.54 = 0.648 \approx 0.6$ (the least accurate value in the original data is 1.2 therefore the answer must be rounded to the nearest tenth).

Example 12.6

1. Round 169 to the nearest ten:
The desired rounding place is 6 (16|9). The digit to the right of 6 is more than 5, therefore round up by increasing 6 by 1 and change all digits to the right of 6 to 0. 169 will become 170.
2. Round 1 819 to the nearest hundred:
The desired rounding place is 8 (18|19). The digit to the right of 8 is less than 5, therefore leave 8 as it is and change all digits to the right of 8 to 0: 1 819 will become 1 800.
3. Round 33.215 to the nearest tenth:
The desired rounding place is 2 (33.2|5). The digit to the right is less than 5, therefore leave 2 as it is and disregard all the digits to the right of 2. 33.215 will become 33.2.
4. Round 5.129 to the nearest hundredth:
The desired rounding place is 2 (5.12|9). The digit to the right of 2 is more than 5, therefore round up by increasing 2 with 1 and change all digits to the right of 2 to 0. 5.129 will become 5.130
5. Round 17.5 to the nearest unit (whole number):
The desired rounding place is 7 and the digit to the right of 7 is the last digit and exactly 5, therefore increase 7 by 1 because 7 is an odd number.
17.5 will become 18

6. Round 19.985 to the nearest hundredth:

The desired rounding place is 8 and the digit to the right of 8 is the last digit and exactly 5, therefore leave 8 as is because 8 is an even number.

19.985 will become 19.98.

Activity 12.16

- A company reports a profit figure last year of R1078245.67. Round this figure to the:
 - nearest million
 - nearest thousand
 - nearest unit
 - nearest tenth
 - Round 539.345 to the nearest hundredth.
 - Round 4.2355 to the nearest thousandth.
 - Round 5.009 to the nearest hundredth.
-

12.15 Significant digits

Significant refers to the number of digits in the number that are accurate – counting from left to right. Rounding can also be done by the number of significant figures we require.

- Start with the non-zero digit (e.g. the '1' in 1 300, or the '2' in 0.0274) on the left.
- Keep the required number of digits and replace the rest with zeros.
- Round up by one if appropriate. For example, if rounding 0.059 to one significant figure, the result would be 0.06.

General rules for determining the number of significant digits in a number:

- All non-zero numbers are significant. For example, the number 163.45 has five significant figures: 1, 6, 3, 4 and 5.
- All zeros between significant numbers are significant, for example the number 5 002 has four significant figures and 301.12 has five significant figures.
- A zero after the decimal point is significant when bounded by significant figures to the left: For example, the number 3 002.0 has five significant figures, 15.2300 has six significant figures, 0.00152300 has six significant figures, 120.00 has five significant figures. If a result accurate to four decimal places is given as 12.23 then it might be understood that only two decimal places of accuracy are available. Stating the result as 12.2300 makes clear that it is accurate to four decimal places.
- The significance of trailing zeros in a number not containing a decimal point can be ambiguous. For example, it may not always be clear if a number like 1 300 is accurate to the nearest unit (and just happens coincidentally to be an exact multiple of a hundred) or if it is only shown to the nearest hundred due to rounding. One method to address this issue is to underline the last significant figure of a number; for example, '80 000' has two significant figures.

- Zeros to the left of a significant figure and not bounded to the left by another significant figure are not significant. For example the number 0.01 only has one significant figure and 0.00012 has two significant figures.
- A number with all zero digits (e.g. 0.000) has no significant digits.

Example 12.7

- Round 742.396 to:
 - 4 significant digits: 742.400
 - 3 significant digits: 742.000
 - 2 significant digits 740.000
- Round 0.06284 to:
 - 4 significant digits: 0.06284
 - 3 significant digits: 0.0628
 - 2 significant digits 0.063
- Round 351.45 to:
 - 4 significant digits: 351.4
 - 3 significant digits: 351
 - 2 significant digits 350
- Round to two significant figures:
 - 13 300 becomes 13 000
 - 14 stays as 14
 - 0.00123 becomes 0.0012
 - 0.4 becomes 0.40 (trailing zero indicates rounding to two significant figures).
 - 0.01084 becomes 0.011
 - 0.0325 becomes 0.032
 - 19 800 becomes 20 000 (see the notes about trailing zeros)

Activity 12.17

Your expense budget for the year amounts to R125 784.66.

The exact budget figure contains eight significant digits. Round this number to:

- one significant digit
 - two significant digits
 - three significant digits
 - six significant digits
 - seven significant digits
-

Round to the appropriate number of significant digits when you add or subtract

Add (or subtract) the numbers as usual, then round the answer to the same decimal place as the least-accurate number.

Example 12.8

1. $13.214 + 234.6 + 7.0350 + 6.38 = 261.2290$

The second number, 234.6, is only accurate to the tenths place, so the answer will have to be rounded to the tenths place:

$$13.214 + 234.6 + 7.0350 + 6.38 = 261.2$$

2. $1\ 247 + 134.5 + 450 + 78 = 1\ 909.5$

450 is only accurate to the tens place, therefore round the final answer to the nearest tens place:

$$1\ 247 + 134.5 + 450 + 78 = 1\ 910$$

Activity 12.18

Calculate the following:

1. $9.812 - 0.13358 + 0.123 =$

2. $1.111 - 0.234 + 0.001 =$

Round to the appropriate number of significant digits when you multiply (or divide)

Multiply (or divide) the numbers as usual, then round the answer to the same number of significant digits as the least-accurate number.

Example 12.9

1. If we multiply 3.3 (rounded to two significant digits) by 3.55 (rounded to three significant digits), the answer is 11.715. This answer appears to have five significant digits; however the result cannot be more accurate than the lowest significant level of the two numbers which is two. The result should be 12.

2. $0.00435 \times 4.6 = 0.02001$

4.6 has only two significant digits, so round 0.02001 to two significant digits.

$$0.00435 \times 4.6 = 0.020$$

(The answer is not 0.02, because this is only one significant digit (the '2'). The trailing zero indicates that 'this is accurate to the thousandth place', and is therefore a necessary part of the answer.

Activity 12.19

Calculate the following:

1. $16.235 \times 0.217 \times 5 =$

2. $0.235 \times 0.0070 \times 1.333 =$

Note: For adding, use 'least accurate place'.

For multiplying, use 'least significant digits'.

12.16 The metric system

This is the most widely used system of weights and measures in the world; a standardised system is necessary for computational purposes in international trade.

The basic units of measurement in the metric system are:

- length, which is measured in metres
- weight, measured in grams
- volume, measured in litres.

The metric system is decimal orientated and deals with powers of 10. Thus, multiplying or dividing by 10 gives the next higher or lower unit.

0.01 centi	0.001 milli	0.1 deci	1 metre 1 litre 1 gram	10 deca	100 hecto	1000 kilo
---------------	----------------	-------------	------------------------------	------------	--------------	--------------

Activity 12.20

Fill in the missing numbers:

kilo	hecto	deca	metre/litre/gram	deci	centi	milli
1000	100	10	1	0.1	0.01	0.001
		6335				
	1250					
			50			
				150		
					1025	
						96

UNIT 13

Percentages and ratios

This unit deals with applying the concept of percentage and ratio calculations in business.

After completion of this unit you will be able to:

- convert percentages to fractions and decimals
- deal with different types of percentage problems
- identify the base
- understand the concept of ratios
- apply percentages and ratios in business.

'Percentage' is derived from the Latin *per centum* meaning 'per hundred' and uses the symbol %. It is a universal basis for comparison whereby a value is expressed as to how much of a hundred such a value represents. The basis of comparison is therefore always 100. Once a value is expressed in terms of its portion of a 100, the result is indicated as a percentage by adding the percentage sign '%' after the result.

13.1 Percentage calculations

Percentages are used widely in business to determine discounts, taxes, interest and numerous comparisons. To use a percentage in an arithmetic application, it must first be changed to a decimal or a fraction.

Commonly used terms:

- Base: the value upon which the percentage is taken.
- Rate (%): the percentage that is taken of the base.

13.1.1 Converting percentages to fractions and decimals

To change a percentage into a fraction or a decimal, divide the number expressing the percent by 100 and drop the % sign. The percent becomes the numerator and the 100 the denominator. The fraction can be converted to a decimal by dividing the numerator by the denominator.

Example 13.1

Express 75% as a fraction and a decimal:

$$75\% = \frac{75}{100} = 0.75$$

Activity 13.1

Find the fraction and decimal equivalents of:

- 44%
- 83%
- 126%

13.1.2 Converting a fraction or decimal into a percentage

To change a fraction or decimal into a percentage, multiply by 100 and add the % sign.

Example 13.2

- Express $\frac{6}{20}$ as a percentage:

$$\frac{6}{20} \times 100 = 30\%$$

- Express 0.14 as a percentage:

$$0.14 \times 100 = 14\%$$

Activity 13.2

- Express $\frac{1}{20}$ as a percentage.
- Express 0.033 as a percentage.

13.1.3 Finding the percentage amount

To find a percentage amount of a value if both the base and the rate are known, you must first convert the rate to a fraction or a decimal number and then multiply by the value, which is the base.

$$\text{percentage amount} = \text{base} \times \frac{\text{rate}}{100}$$

Example 13.3

Calculate 5% of R200:

$$\frac{5}{100} \times \text{R}200 = \text{R}10$$

Activity 13.3

1. Calculate 14% of 430.
 2. Calculate 5% of 684.
-

13.1.4 Finding the rate

To determine the percentage rate when both the base and the percentage amount are known, you must first construct a fraction using the percentage amount as numerator and the base as the denominator.

$$\text{rate (\%)} = \frac{\text{percentage amount}}{\text{base}} \times 100$$

Example 13.4

What % is 8 of 18?

$$\frac{8}{18} \times 100 = 44.44\%$$

Activity 13.4

1. What percentage is 9 of 18?
 2. What percentage is 3 of 67?
-

13.1.5 Finding the base

If the rate and the percentage amount are known, the base can be determined by dividing the percentage amount by the rate.

$$\text{base} = \frac{\text{percentage amount}}{\text{rate}} \times 100$$

Example 13.5

If 14% of a value equals 90, what is the value?

$$\frac{90}{14} \times 100 = 642.86$$

Activity 13.5

If you receive R4 600 simple interest on an investment earning 9%, how much did you invest?

Hidden page

Activity 13.7

1. The total weight of a packaged article is 25% greater than its nett weight of 6 kg. Determine the total packaged weight.
 2. If an article weighs 9 kg after it is packaged, and the increase in weight is 20%, how much did the article weigh before packaging?
-

13.2 Ratio (proportion) calculation

A fraction and a ratio are different ways of expressing the same relationship. A ratio is a comparison of two quantities expressed in the same measurement units and may be written as follows: a to b or $a:b$ or $\frac{a}{b}$.

Ratios are also used to express rates such as 120 kilometres/hour, 2.3 children/family, 14 kilometres/litre of petrol.

Example 13.8

1. In comparing two drums of paint, one with a volume of 20 litres and the other a volume of 50 litres, we can say that the ratio between the two drums is 20:50 or 2:5. That means that the small drum has $\frac{2}{5}$ the volume of the big drum, which is also $\frac{2}{5} \times 100 = 40\%$. Alternatively we can say that the volume of the big drum is $\frac{5}{2} = 2.5$ times more than the small drum or $2.5 \times 100 = 250\%$ more.
2. If examinations normally result in a failure rate of seven per 200 students, the number of failures that can be expected if 800 students write the examination is: $\frac{7}{200} \times 800 = 28$ students.

Activity 13.8

1. Three friends decided to contribute R20, R10 and R5 respectively to buy tickets from the national lottery. Their agreed division of winnings will be in the same ratio as their contributions. If their winnings amount to R500, what sum of money will each one receive?
 2. A golf player receives 10% of his income from sponsorships, three times as much from training, and the rest from tournaments. If his total income for the year is R340 000, how much did he get from each source?
-

13.3 Business applications

Although percentages have many applications in many disciplines, in the manufacturing, retail and wholesale environment they are usually applied to pricing of goods or services, to determine final prices after adding profit margins or allowing for discounts. Percentages can also be used in stock control levels.

The cost price is the price a wholesaler or retailer paid for a product or service excluding the VAT, or what the cost was to manufacture the product from scratch by the manufacturer.

The 'selling price' is the price for which a product or service is sold.

13.3.1 Mark-up on cost price

The manufacturer sells his product to a retailer, wholesaler or final consumer and will want to do so at a profit. A 'mark-up' margin must therefore be added to the product to determine the 'selling price' of the item.

The mark-up margin is usually expressed as a percentage to be added to the cost price.

Example 13.9

ABS Manufacturers produce classroom desks used in schools. The cost price of a desk is R120.00 and ABS adds a mark-up of 40% to this price to determine the selling price per unit.

The selling price per unit is therefore: $R120.00 + \left(\frac{40}{100} \times 120\right) = R168.00$

ABS will now sell these tables to Tablecor (Pty) Ltd at R168.00 + VAT per desk. A desk will cost Tablecor $R168.00 + \left(\frac{14}{100} \times 168\right) = R191.52$

Assume that Tablecor decides to add its own profit or mark-up. Tablecor now sells the desks to the Education Department for use in schools. The cost price of a desk for Tablecor is R168.00 (because they reclaim the VAT that they have paid as input VAT).

Tablecor uses a mark-up percentage of 20%.

The price at which Tablecor will then sell each desk is:

$R168.00 + \left(\frac{20}{100} \times 168\right) = R201.60$, excluding VAT. Adding VAT to the selling price means that the customer will pay:

$R201.60 + \left(\frac{14}{100} \times 201.60\right) = R229.82$ per desk.

Activity 13.9

Calculate the following without taking VAT into account:

1. An article costs R32. It is sold at a profit of 15%. Find the selling price.
 2. By selling an article for R63.50, a profit of 25% is made. What is the cost price?
 3. An article costs R25 and is sold at a profit of R3. What is the % profit?
 4. An article costs R250 and is sold for R300. What is the % profit?
-

13.3.2 Mark-downs and discounts

While a mark-up means adding to a base price, markdown means reducing a base price. A mark-down differs from a discount in the sense that a discount is a reduction in selling price because of method of payment (cash) or due to volumes purchased or

Hidden page

Hidden page

TEST YOURSELF 13

1. At the end of 2001 there were 101 stores open in South Africa – 39 in Gauteng, 33 in Cape Town, 19 in Natal and 10 in the Free State. Find the percentage of each in relation to the total.
2. Each section in a department store is given a target for the year, with Jack's section targeted for an increase of 25% over last year's results. If last year's sales were R1.5 million, what was Jack's targeted sum?
3. A salesman's commission makes up 13% of his total weekly income. If his commission is R948 for a particular week, what is his total income?
4. In the past 10 years, employment in a company has fallen by 780 to 3 240. What is the percentage decline in employment over the decade?
5. Your commission of R475 for the week is equivalent to what rate of commission if the commission is calculated on weekly sales of R5 000?
6. With a trade discount of 30%, an electrician paid R83.30 for materials. What would the material have cost without the trade discount?
7. Mr Hammer, a carpenter, receives a trade discount of 20% and a cash discount of 3% at the local hardware store. The list price for materials he purchased totalled R4 532.50. Find the invoice price and the actual amount paid cash for the material. What percentage of the list price did Mr Hammer save?
8. Giftware often carries a mark-up cost of 50%. If the cost of a vase to a retailer is R84, what will the vase retail for?
9. A pair of shoes priced at R793 has been marked-up on cost by 30%. What was the cost price to the retailer?
10. The cost to the retailer of a tennis racquet that sold for R165 was R110. Find the profit as a percentage based on cost, and the profit percentage based on the list price.
11. The cost of a garment to a boutique was R400. If a profit margin of 17% was made on the retail price, find the retail price. If a loss of 2.5% was made on the retail price, what was the retail price?
12. All candles in a particular gift shop are priced at R14.25 after a mark-up on cost of 25%. What was the cost of the candle to the proprietor?
13. A dealer marks all baby goods 25% above cost. What percentage discount can he allow during a sale to ensure a profit of at least 10%.
14. Goods are bought for R30. At what price must they be marked in order to yield 10% profit after a trade discount of 10% has been allowed?
15. A suit is marked at R999.99. A trader allows 2.5% discount and still makes 20% profit. What did the suit cost the trader?
16. You purchase a Hi-Fi set and a DVD player for R3 400 and R2 800 respectively. How much VAT in total will you have to pay on these transactions?
17. Sipho buys a television set from Grand Bazaars for R4 600. Because he pays cash for it, he is given a discount of 15% on the purchase.
 - a) How much does Sipho pay for the TV before VAT is charged?
 - b) How much discount did Sipho get?
 - c) How much VAT is charged on this transaction?
 - d) What is the final amount that Sipho pays for his TV set?

18. Smart City Appliances receives a delivery of 25 refrigerators from Freezer Manufacturers together with an invoice for R102 600. Upon inspection, Smart City returns four refrigerators because they are damaged, and requests a credit for the returned goods. Freezer has also granted Smart City 10% trade discount on the order. The discount has already been included in the invoice.
- How much did a refrigerator originally cost?
 - How much discount did Smart City get per freezer?
 - How much VAT is included in the original invoice?
 - How much credit must Smart City get, excluding VAT?
 - By how much must Freezer adjust the VAT charged?
 - How much will Smart City now have to pay Freezer Manufacturers?
19. Smart Stores sells fashion clothing directly to the public. The prices of all items are inclusive of VAT. The price reflected on the price tag of a garment is the price to be paid. Agnes purchases three dresses at R160.00 each and a track suit for R210.00.
- How much does she have to pay Smart Stores for the purchases?
 - What were the prices of the dresses and track suit before Smart Stores added the VAT?
 - How much VAT did she pay on the whole transaction?
 - How much will Agnes have to pay if Smart Stores grants her 10% discount on the dresses and 8% discount on the track suit?
 - How much VAT will Smart Stores add to the transaction if they grant the discounts above?
20. Eight slabs of chocolate cost R32. Find the cost of three slabs of chocolate.
21. John takes 30 minutes to walk from his home to school at a speed of 4 km/hr. how long will he take if he cycles at 10km/hr?
22. A lecturer takes three hours to mark the books of all the students in her class. How long will it take three lectures to mark the same books if they work at the same pace?
23. It takes three markers 120 hours to mark the students examination scripts. Assuming they all work at the same pace, calculate how long it will take if there are:
- six markers
 - 10 markers
 - 20 markers.
24. If I travel at 50 km/hr I can do a journey in six hours. How long will it take the same journey at 40 km/hr?
25. A farmer buys enough chicken feed to last 200 chickens for a week. How long will the same amount of feed last for 350 chickens? (Each fowl eats the same amount each day.)
-

UNIT 14

Equations and graph construction

In this unit we look at solving equations and ways to make this easier.

After completion of this unit you will be:

- familiar with the concept of graph construction
- familiar with business applications using linear equations.

One of the most important concepts of mathematics concerns the relationship between the elements of sets. In this unit we are mainly concerned with relationships between two sets of numbers – such as numbers representing supply and demand; age and value of machinery; unit cost and the number of units produced and so on.

14.1 Graph construction

A graph shows a picture of the trend or relationship between two variables (x and y) – that is, how one quantity changes with respect to another.

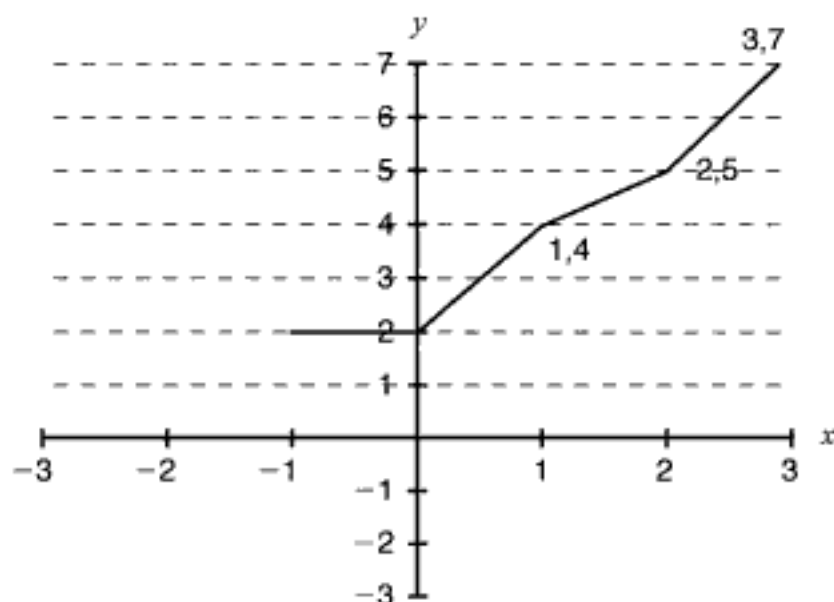
The type of graph to be drawn depends on the type of data, the complexity of the data and the requirements of the user. In this text, we will deal with the linear graph only.

Two variable functions are graphed on a set of rectangular coordinate axes. The plane formed by the coordinate axes is called the *Cartesian plane*. In order to set up a Cartesian graph, the following steps must be followed:

- Two lines, known as coordinate axes, are drawn at right angles dividing the plane into four quadrants. The point where the two lines cross is known as the origin (0). The horizontal line is known as the x -axis and the vertical line as the y -axis.
- Indicate units of length or a scale on the two axes (not necessarily the same for each one). To select a scale, determine the maximum and minimum numbers you will use for each variable and subdivide the axis in multiples of, for example, 1, 2, 3, 5, 10 and 100 as necessary to accommodate the maximum and minimum number of each variable. To the right of the y -axis, x is positive and to the left, it is negative. Above the x -axis, y is positive and below, it is negative.
- All values along the x -axis are known as *abscissas* and are plotted below the x -axis.
- All the values along the y -axis are known as *ordinates* and are shown to the left of the vertical axis.
- The graph should have a title and both the axes should be labelled.
- Any point in the Cartesian plane is defined by an ordered pair of coordinates (x, y), with the value of x always given first.

- A mathematical function assigns one value of y to each value of x within its equation and by arbitrarily selecting values for x , a corresponding value for y can be computed with a resulting set of ordered paired coordinates (x, y) .
- Each of these pairs of coordinates corresponds to a point on the Cartesian plane and if we plot all the points we obtain the graph of the function.

For example, the coordinate point $(1, 4)$ is exactly 1 unit to the right of the 0 along the horizontal line and 4 units above the zero on the vertical line.



Note: The Cartesian plane has four quadrants. While all are mathematically important, we find that in business the most important of the four is the top right quadrant where both x and y have positive values.

14.2 Solution of equations

An equation is a shorthand way of stating that two algebraic expressions are equal. It has a left hand side that is equal to a right hand side. To solve an equation, the value of the unknown (variable) must be calculated. If it is substituted into the equation, the value on the left must equal the value on the right.

14.2.1 Linear equations

Any mathematical function that appears as a straight line when plotted in the coordinate plane is a linear function. One of the most commonly used equations in business is:

$$y = a + bx$$

Where $a = y$ -intercept – that is the point on the y -axis where the line will cut
 $b =$ slope or gradient.

The slope can be measured between any two points on the line – it is always the same – and can be defined as the: $\frac{\text{increase in } y}{\text{increase in } x}$

You can interpret it as follows: it is the number of units the line rises or falls vertically (y -axis) for each unit of horizontal (x) change from left to right.

When the slope (b) is positive the line has an increasing trend and if b is negative the line has a decreasing trend. In business the slope is seen as the ratio of change in y to the change in x or the marginal value.

Some examples of linear functions to measure profitability in business are:

- the linear cost function
- the linear income function
- the linear profit function.

14.2.2 Linear cost function: $C(x)$

Organisations are concerned about costs because they reflect money flowing out of the business. These costs are usually to pay for salaries, raw materials, rent, municipal charges and so forth.

Cost is defined in terms of two components: total variable cost and total fixed cost. These two components must be added to obtain the total cost. Variable costs vary with the level of output. The linear cost function is:

$$C = F + Vx$$

Where:

C = total cost

V = variable cost per unit

F = fixed cost per period

Example 14.1

A company which produces a single product wants to determine the function that expresses total cost (C) as a function of the number of units produced (x). The fixed expenditure each year is R50 000. The estimated raw material and labour cost for each unit produced is R5.50. What will the total cost be to produce 120 items?

$$\begin{aligned} C(x) &= 50\,000 + 5.50x \\ C(x=120) &= 50\,000 + 5.50(120) \\ &= 50\,660 \text{ rand} \end{aligned}$$

The y -intercept tells us that the cost of producing zero units is R50 000. This is the fixed cost. The slope tells us that for each unit that the line moves to the right, the cost increase by R5.50. Therefore, the cost of producing one extra unit each time is R5.50 and this is then the marginal cost of the product.

Activity 14.1

1. Peter is setting up a small home business to manufacture an item he has developed. He has invested R10 000 in equipment and can produce each item for R0.65. Determine the cost function for Peter's product.

2. A car rental agency leases cars at a rate of R100 per day plus R2.00 per kilometre driven. Determine the function which expresses the daily cost of renting a car as a function of the number of kilometres driven in one day. What will the total cost be for 315 km?
3. The police department is contemplating the purchase of an additional patrol car. Police analysts estimate the purchase cost of a fully equipped car to be R180 000. They have also estimated an average operating cost of R4.00 per kilometre. Determine the function that represents the total cost of owning and operating the car in terms of the number of kilometres it is driven. What are the projected costs if the car is driven 50 000 km during its lifetime?

14.2.3 Linear revenue function

The money that flows into a business from either selling products or providing services is referred to as revenue. If we assume that the selling price is the same for all units sold, then

$$\text{total revenue (R)} = \text{price (p)} \times \text{quantity (x)}$$

Example 14.2

A local car rental agent is trying to compete with some of the larger companies and bought good second hand cars for his fleet. He also simplified the rental rate structure by charging a flat R125 per day for the use of the car. The total linear revenue function is $R = 125x$

If a car was rented out for 20 days last month, what was the total revenue for the car?
 $R(x = 20) = 125(20) = 2\,500$ rand

14.2.4 Linear profit function

Profit is the difference between total revenue and total cost.

$$P(x) = R(x) - C(x)$$

When total revenue exceeds total costs, profit is positive and is referred to as nett gain. When total costs exceed total revenue, profit is negative and it is called nett loss or deficit.

Example 14.3

The price of a single product is R65. Variable costs per unit are R20 for materials and R27.50 for labour. Annual fixed costs are R100 000. Construct the profit function and determine the profit if annual sales are 20 000 units.

$$C(x) = 100\,000 + 47.50x$$

$$R(x) = 65(x)$$

$$\begin{aligned} P(x) &= 65(x) - (100\,000 + 47.50x) \\ &= -100\,000 + 17.50x \end{aligned}$$

If 20 000 units are sold, the profit will be:

$$\begin{aligned} P_{(20\,000)} &= -100\,000 + 17.50(20\,000) \\ &= 250\,000 \text{ rand} \end{aligned}$$

14.2.5 Break-even analysis

Break-even analysis is used to determine the number of units that must be sold (either in rands or units of output) for the business to break even; that is, to neither earn profits nor incur losses. The break-even point will be achieved when sales produce just enough revenue above variable costs to cover fixed costs.

Steps

1. Construct the total cost function $C(x)$ where x represents the level of output.
2. Construct the total revenue function $R(x)$.
3. Set $C(x) = R(x)$ and solve x .

Example 14.4

A product is priced at R10 and the variable cost is R6 per unit. If total fixed costs are R1 000, the breakeven point in units of output sold is:

$$C(x) = 1\,000 + 6x \qquad R(x) = 10x$$

$$10x = 1\,000 + 6x$$

$$10x - 6x = 1\,000$$

$$4x = 1\,000$$

$$x = 250 \text{ units}$$

250 units at R10 each will give a break-even income of R2 500.

TEST YOURSELF 14

1. An engineer is interested in forming a company to produce smoke detectors. The estimated variable costs per unit, including materials and labour, are R22.50. Fixed costs associated with the formation, operation and management of the company, as well as the purchase of equipment and machinery, total R250 000. A market related selling price would be R30 per detector.
 - a) Determine the number of smoke detectors that must be sold in order for the engineer to break even.
 - b) Determine the break-even value.
 - c) If marketing research indicated that the firm can expect to sell approximately 30 000 smoke detectors over the life of the project, determine expected profits at this level of output.

2. A company produces a product which sells at a price of R25 per unit. Variable costs are estimated to be R18.75 per unit and fixed costs are R50 000.
 - a) Determine the break-even level of output.
 - b) Compute the total cost and total revenue at the break-even point.
 - c) What will profit equal if demand equals 7 500 units?
 3. A local Gauteng charity organisation is planning a one week holiday in Cape Town. The venture is a fund-raising effort. A package deal has been worked out with a commercial airline whereby the charity will be charged a fixed cost of R10 000 plus R300 per person. The R300 covers the flight cost, airport tax, hotel and meals. The organisation is planning to price the package at R450 per person.
 - a) Determine the number of persons necessary to break even on the venture.
 - b) The goal of the organisation is to net a profit of R10 000. How many people must participate for the goal to be realised?
-

Hidden page

available before the end of the term and the interest is not added to the principle to earn interest on interest.

The standard formulae for calculating simple interest are:

$$\begin{aligned} I &= Prt \\ A &= P(1 + rt) \\ A &= P + I \end{aligned}$$

Where:

I = amount of interest

P = principal

A = amount

r = interest rate per annum expressed as a decimal

t = time in years or a portion of a year

Note: Exact interest is calculated on a basis of 365 days per year or 366 in a leap year. Ordinary interest is calculated on a basis of 360 days per year or 30 days per month.

Example 15.1

1. Thandi borrows R5 000 from Simon. Thandi must repay the R5 000 before the end of 12 months and the interest is 15% per year.

How much must Thandi pay Simon after 12 months?

$$\begin{aligned} I &= Prt \\ &= 5\,000(0.15)(1) \\ &= 750 \end{aligned}$$

$$\begin{aligned} A &= P + I \\ &= 5\,000 + 750 \\ &= 5\,750 \text{ rand} \end{aligned}$$

2. Determine the present value at 15% simple rate of interest on an amount of R12 500 due in one year and nine months.

$$\begin{aligned} A &= P(1 + rt) \\ 12\,500 &= P[1 + (0.15 \times 1.75)] \\ P &= R9\,900.99 \text{ rand} \end{aligned}$$

3. B borrows R500 from A and at the end of eight months pays A an amount of R525. What is the simple rate of interest earned?

$$\begin{aligned} I &= Prt \\ 25 &= 500(r)\left(\frac{8}{12}\right) \\ r &= 0.075 \text{ that is } 7.5\% \end{aligned}$$

4. How long will it take R5 000 to earn R50 interest at 10%?

$$\begin{aligned} I &= Prt \\ 50 &= 5\,000(0.10)t \\ t &= 0.10 \text{ of a year which is one month and six days} \end{aligned}$$

Hidden page

Example 15.2

1. Simon loans R1 000 to Thandi at a rate of 15% per annum calculated monthly. What is the amount she must repay at the end of two years?

The interest rate of 15% is the interest that is charged for the year. However if the interest is to be calculated monthly, then the annual interest rate (15%) must be converted to a monthly interest rate by dividing by 12:

$$i = \frac{15\%}{12} = 1.25\%$$

The two year time period should change to $n = 12 \times 2 = 24$

$$A = 1\,000\left(1 + \frac{1.25}{100}\right)^{24} = 1\,000(1.0125)^{24} = \text{R}1\,347.35$$

Amount of interest paid:

$$\text{R}1\,347.35 - \text{R}1\,000 = \text{R}347.35$$

2. A young man inherited R200 000. He wants to invest a portion of his inheritance to accumulate R300 000 in 15 years. What portion of the money should be invested if the money will earn 8% per year compounded semi-annually and how much interest will be earned over the period?

$$i = 4\% \quad n = 30$$

$$A = P(1 + i)^n$$

$$300\,000 = P\left(1 + \frac{4}{100}\right)^{30}$$

$$P = \frac{300\,000}{3.2434}$$

$$= 92\,495.53 \text{ rand}$$

$$I = 300\,000 - 92\,495.53 = 207\,504.47 \text{ rand}$$

3. Determine the interest rate on a study loan which would increase its value from R36 000 to R50 000 in five years if the interest is compounded monthly.

$$i = \left(\frac{A}{P}\right)^{\frac{1}{n}} - 1$$

$$= \left(\frac{50\,000}{36\,000}\right)^{\frac{1}{60}} - 1$$

$$= 0.0055$$

Monthly rate is $0.0055 \times 12 \times 100 = 6.6\%$

4. How long will it take for R20 to amount to R30 at 5% compounded quarterly?

$$t = \frac{\log \frac{A}{P}}{\log(1 + i)}$$

$$= \frac{\log \frac{30}{20}}{\log\left(1 + \frac{1.25}{100}\right)}$$

$$= 32.64 \text{ quarters} \approx 8 \text{ years, 1 month, 28 days}$$

Activity 15.2

1. Find the present value of R2 000 due in 18 months if money is worth 11% compounded semi-annually.

Hidden page

Annuities are classified into two main classes:

- **Ordinary annuities certain** refer to annuities where the regular payments are made at the end of each payment interval.
- **Ordinary annuities due** refer to annuities where the periodic payment (R) falls at the beginning of each payment interval.

15.5.1 Ordinary annuities certain

The regular payments are made at the end of each payment period.

To calculate the future value or amount (A) of an ordinary annuity certain we apply the following formula:

$$A = R \frac{(1+i)^n - 1}{i}$$

To calculate the present value or principle (P) of an ordinary annuity certain we apply the following formula:

$$P = R \left(\frac{1 - (1+i)^{-n}}{i} \right)$$

Example 15.4

1. Determine the amount of an annuity certain of R150 per month for three years if money is worth 12% compounded quarterly.

$$\begin{aligned} A &= R \frac{(1+i)^n - 1}{i} \\ &= 150 \frac{(1 + \frac{3}{100})^{12} - 1}{\frac{3}{100}} = 2\,128.80 \text{ rand} \end{aligned}$$

2. A student needs R3 000 a year for books for four years with the first R3 000 available one year from now. If the student can get 8% p.a. return on investment, how much money should he invest now?

$$\begin{aligned} P &= R \left(\frac{1 - (1+i)^{-n}}{i} \right) \\ &= 3\,000 \left(\frac{1 - (1 + \frac{8}{100})^{-4}}{\frac{8}{100}} \right) = 9\,936.38 \text{ rand} \end{aligned}$$

3. Arthur wants to have R6 000 in the bank in five years' time. He plans to deposit the correct amount to achieve this, at the end of each month. What should the value of each monthly payment be if interest is 15% compounded monthly?

$$\begin{aligned} A &= R \frac{(1+i)^n - 1}{i} \\ \therefore 6\,000 &= R \left(\frac{(1 + \frac{1.25}{100})^{60} - 1}{\frac{1.25}{100}} \right) \\ \therefore R &= \frac{6\,000}{88.5745} \\ &= 67.74 \text{ rand} \end{aligned}$$

Hidden page

2. The premium on a life insurance policy is R60 per quarter, payable in advance. Determine the cash equivalent of a year's premiums if the insurance company charges 10% compounded quarterly for the privilege of paying this way instead of all at once for the year?

$$P = R \left(\frac{[1 - (1 + i)^{-n}][1 + i]}{i} \right)$$

$$= 60 \left(\frac{[1 - (1 + \frac{2.5}{100})^{-4}][1 + \frac{2.5}{100}]}{\frac{2.5}{100}} \right) = 231.36$$

3. The beneficiary of a life insurance policy may take R10 000 in cash or 10 equal payments, the first to be made immediately. What is the annual payment if money is worth 12%?

$$P = R \left(\frac{[1 - (1 + i)^{-n}][1 + i]}{i} \right)$$

$$10\,000 = R \left(\frac{[1 - (1 + \frac{12}{100})^{-10}][1 + \frac{12}{100}]}{\frac{12}{100}} \right)$$

$$\therefore R = \frac{10\,000}{6.3282} = 1\,580.23$$

4. The Bell Company plans to open a new retail outlet in its chain of telephone equipment stores three years from today. How much must Bell invest at the beginning of each semi-annual to have enough for the estimated costs of R100 000, if the interest rate is 9%?

$$P = R \left(\frac{[1 - (1 + i)^{-n}][1 + i]}{i} \right)$$

$$100\,000 = R \left(\frac{[(1 + \frac{4.5}{100})^6 - 1][1 + \frac{4.5}{100}]}{\frac{4.5}{100}} \right)$$

$$\therefore R = \frac{100\,000}{7.0192} = 14\,246.64$$

Activity 15.5

- The rent of a building is R15 000 per year payable in advance. If the interest rate is 6% compounded monthly, what will the equivalent monthly rental, payable in advance, be?
- Mr Cute bought a car paying R2 000 deposit and R200 at the beginning of each week for two years. If the interest rate is 9% compounded weekly, what was the cash price of the car?
- A school sets aside R10 000 at the beginning of each year to create a fund in case of further expansion. If the fund earns 5%, how much does it amount to at the end of the seventh year?
- A debt of R5 000, inclusive of 5% interest compounded quarterly, is to be settled within three years in equal quarterly payments. If the first payment is due today, what will be the size of each payment?

TEST YOURSELF 15

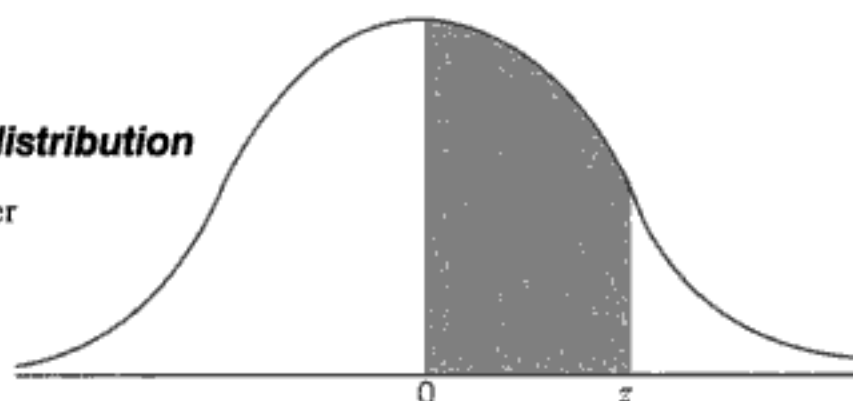
1. Using the simple interest approach, how much interest will you pay on a loan of R15 000 at 12.5% interest per annum? How much must you repay after three years and six months?
2. If the amount plus interest that George had to repay at the end of a one year loan was R11 000 and the interest rate was 10% per annum, what was the amount of the principal sum? Use the simple interest formula to calculate.
3. A waitress who was temporarily pressed for funds pawned her watch and diamond ring for R55. At the end of one month she redeemed them by paying R59.40. What was the annual rate of interest?
4. A 26% interest charge on an overdue account of R800 came to R21. How late was the account?
5. A mechanic borrowed R125 from a cash loan company and at the end of one month paid off the loan with R128.75. What annual rate of interest was paid?
6. At what rate will simple interest on R1 127 amount to R318 in 135 days?
7. John invests R200 at 7.75% simple interest per annum and receives R295 after a certain time. For how long was the money invested?
8. Philemon has an option of financing the purchase of a new music centre with a price of R800, through a loan for one year. The interest rate on the loan is 12% per annum. He has an option of taking a loan with interest calculated quarterly or a loan where the interest is calculated semi-annually. Which option will you recommend to Philemon? The lender will apply the compound interest formula.
9. The outstanding amount on your account is R2 650. If the store charges 24% interest compounded monthly, how much will you owe after three months if no payment was made during that period?
10. A cell phone company will need R500 000 to replace a piece of equipment in eight years. How much must be invested now at 6% compounded quarterly to accumulate this amount?
11. If R500 amounts to R700 in five years with interest compounded quarterly, what is the rate of interest?
12. A cash loan company charges 36% compounded monthly on small loans. How long will the loan company take to triple its money at this rate?
13. How long will it take R4 000 to amount to R5 000 at 9% compounded quarterly?
14. What is the effective rate of interest equivalent to 15% converted semi-annually, quarterly and monthly?
15. What is the nominal rate of interest, compounded semi-annually and monthly, equivalent to 24% effective?
16. Which gives the better annual return on investment, 4% compounded quarterly, 4% converted semi-annually or 4% converted monthly?
17. A refrigerator can be bought for R50 deposit and R28 per month for 24 months, payable at the end of each month. What is the equivalent cash price if the rate is 26%?

18. A company bought a machine costing R8 000 and estimates that its useful life will be five years, after which it will be sold as scrap for R300. The company decides to set up a reserve fund to cover the cost of a replacement machine in five years time. Equal amounts are to be invested at the end of each year in an account that earns 10% compound interest. Due to inflation, it is estimated that the cost of this machine will be R15 000. How much must be invested each year to cover the cost of a replacement machine, allowing for the scrap value of the present one?
 19. If money is worth 15% compounded quarterly, what single payment today is equivalent to 15 quarterly payments of R100 each, the first due three months from today?
 20. A cash loan company charges 36% converted monthly for small loans. What would be the payment at the end of every month if a loan of R250 is to be repaid within one year?
 21. Mr Smith invests R20 at the end of every week at 18% compounded weekly. What amount will be in his savings account after six months?
 22. A student wants to save R15 000 for a trip after graduation, four years from now. How much must she save at the end of every six months if she gets 15% compounded semi-annually?
 23. Mr T. Bone took out a R100 000 loan on a steakhouse over a 10-year period at an interest rate of 12% compounded monthly. After 3.5 years, interest rates climbed to 15% compounded monthly. If his repayments were at the end of each month, how much did Mr. Bone owe at the end of the first 3.5 years? What was his monthly repayment for the remaining 6.5 years?
 24. Instead of taking R5 000 from an inheritance Peter decides to take monthly payments for a period of five years, with the first to be made immediately. If interest is 6% compounded monthly, what will be the size of each payment?
 25. Instead of paying R1 250 rent at the beginning of each month for the next eight years, Mary decides to buy a flat. Considering interest of 15% to be compounded monthly, what is the cash equivalent of the eight years' rent?
 26. At the beginning of each semester, Abdul invests R900 at an interest rate of 7% compounded semi-annually to guarantee a sum sufficient to start a practice for his daughter, who is entering medical school. If his daughter finishes within eight years, how much will Abdul have for the practice?
 27. Dr Kaye wants to spend five years researching a new book on the motor industry. He calculated that he needs R12 000 a month to live on over the five years. How much must Dr Kaye deposit today in an account earning 12% interest compounded monthly in order to withdraw R12 000 at the beginning of each month?
 28. James wants to accumulate R500 within the next three months by depositing money in a savings account at the beginning of each week. The bank pays 4% compounded weekly. How much must he deposit every week to reach his target?
 29. An investment of R20 is made at the beginning of each day in the money market for one year at 6% compounded daily. How much will the investment be worth at the end of the year?
-

Appendix 1

The standard normal distribution

This table gives the area under the standard normal curve between 0 and z , i.e. $P[0 < Z < z]$

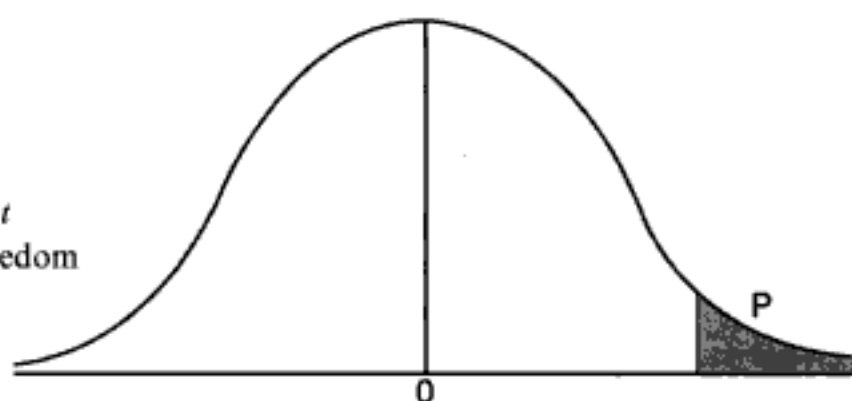


Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1027	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2703	0.2734	0.2764	0.2793	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.48928	0.48956	0.48983	0.49010	0.49036	0.49061	0.49086	0.49111	0.49134	0.49158
2.4	0.49180	0.49202	0.49224	0.49245	0.49266	0.49286	0.49305	0.49324	0.49343	0.49361
2.5	0.49379	0.49396	0.49413	0.49430	0.49446	0.49461	0.49477	0.49492	0.49506	0.49520
2.6	0.49534	0.49547	0.49560	0.49573	0.49585	0.49598	0.49609	0.49621	0.49632	0.49643
2.7	0.49653	0.49664	0.49674	0.49683	0.49693	0.49702	0.49711	0.49720	0.49728	0.49736
2.8	0.49744	0.49752	0.49760	0.49767	0.49774	0.49781	0.49788	0.49795	0.49801	0.49807
2.9	0.49813	0.49819	0.49825	0.49831	0.49837	0.49841	0.49846	0.49851	0.49856	0.49861
3.0	0.49865	0.49869	0.49874	0.49878	0.49882	0.49886	0.49889	0.49893	0.49897	0.49900
3.1	0.49903	0.49906	0.49910	0.49913	0.49916	0.49918	0.49921	0.49924	0.49926	0.49929
3.2	0.49931	0.49934	0.49936	0.49938	0.49940	0.49942	0.49944	0.49946	0.49948	0.49950
3.3	0.49952	0.49953	0.49955	0.49957	0.49958	0.49960	0.49961	0.49962	0.49964	0.49965
3.4	0.49966	0.49968	0.49969	0.49970	0.49971	0.49972	0.49973	0.49974	0.49975	0.49976
3.5	0.49977	0.49978	0.49978	0.49979	0.49980	0.49981	0.49981	0.49982	0.49983	0.49983
3.6	0.49943	0.49985	0.49985	0.49986	0.49986	0.49987	0.49987	0.49988	0.49988	0.49989
3.7	0.49989	0.49990	0.49990	0.49990	0.49991	0.49991	0.49991	0.49992	0.49992	0.49992
3.8	0.49993	0.49993	0.49993	0.49994	0.49994	0.49994	0.49994	0.49995	0.49995	0.49995
3.9	0.49995	0.49995	0.49996	0.49996	0.49996	0.49996	0.49996	0.49996	0.49997	0.49997
4.0	0.49997	0.49997	0.49997	0.49997	0.49997	0.49997	0.49998	0.49998	0.49998	0.49998

Appendix 2

The *t*-distribution

This table gives the value of *t*
where *df* is the degrees of freedom

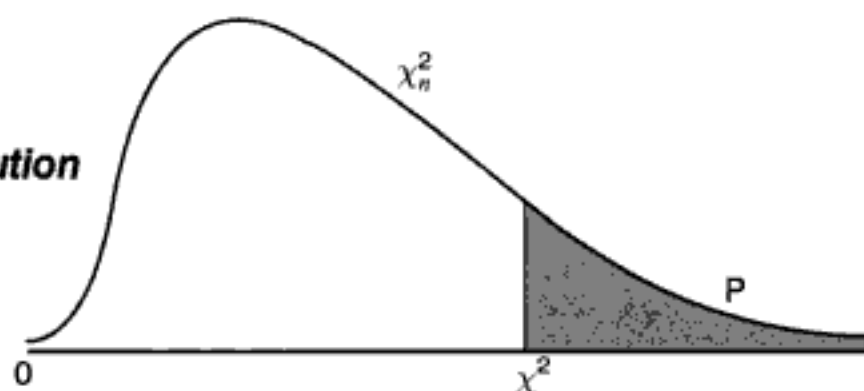


P	ONE-TAIL TEST								
	0.200	0.100	0.050	0.025	0.010	0.005	0.0025	0.0010	0.0005
	0.40	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
DF									
1	1.376	3.078	6.314	12.706	31.821	63.657	127.322	318.313	636.633
2	1.061	1.886	2.920	4.303	6.965	9.925	14.089	22.237	31.599
3	0.978	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.941	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.160
5	0.920	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.906	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.896	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.889	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.883	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.879	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.876	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.873	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.870	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.868	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.866	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.865	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.863	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.862	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.861	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.860	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.859	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.858	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.858	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.857	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.856	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.856	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.855	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.855	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.854	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.854	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
31	0.853	1.309	1.696	2.040	2.453	2.744	3.022	3.375	3.633
32	0.853	1.309	1.694	2.037	2.449	2.738	3.015	3.365	3.622
33	0.853	1.308	1.692	2.035	2.445	2.733	3.008	3.356	3.611
34	0.852	1.307	1.691	2.032	2.441	2.728	3.002	3.348	3.601
35	0.852	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591
36	0.852	1.306	1.688	2.028	2.434	2.719	2.990	3.333	3.582
37	0.851	1.305	1.687	2.026	2.431	2.715	2.985	3.326	3.574
38	0.851	1.304	1.686	2.024	2.429	2.712	2.980	3.319	3.566
39	0.851	1.304	1.685	2.023	2.426	2.708	2.976	3.313	3.558
40	0.851	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
45	0.850	1.301	1.679	2.014	2.412	2.690	2.952	3.282	3.520
50	0.849	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	0.848	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
70	0.847	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435
80	0.846	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
90	0.846	1.291	1.662	1.987	2.369	2.632	2.878	3.183	3.402
100	0.845	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.391
110	0.845	1.289	1.659	1.982	2.361	2.621	2.865	3.166	3.381
120	0.845	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.374
140	0.844	1.288	1.656	1.977	2.353	2.611	2.852	3.150	3.361
160	0.844	1.287	1.654	1.975	2.350	2.607	2.847	3.142	3.352
180	0.844	1.286	1.653	1.973	2.347	2.603	2.842	3.136	3.346
200	0.843	1.286	1.653	1.972	2.345	2.601	2.839	3.132	3.340
∞	0.841	1.282	1.645	1.960	2.327	2.576	2.807	3.091	3.291

Appendix 3

The chi-square distribution

Upper percentage points ($P < 0.05$)



P	0.200	0.100	0.050	0.025	0.01	0.005	0.0025	0.0010	0.0005
DF									
1	1.643	2.707	3.843	5.026	6.637	7.881	9.142	10.829	12.117
2	3.219	4.605	5.991	7.378	9.210	10.597	11.983	13.816	15.202
3	4.642	6.251	7.815	9.348	11.345	12.838	14.321	16.267	17.731
4	5.989	7.779	9.488	11.143	13.277	14.860	16.424	18.467	19.997
5	7.289	9.236	11.071	12.833	15.086	16.750	18.386	20.515	22.105
6	8.558	10.645	12.592	14.449	16.812	18.548	20.249	22.458	24.103
7	9.803	12.017	14.067	16.013	18.475	20.278	22.040	24.322	26.018
8	11.030	13.362	15.507	17.535	20.090	21.955	23.774	26.124	27.868
9	12.242	14.684	16.919	19.023	21.666	23.589	25.462	27.877	29.666
10	13.442	15.987	18.307	20.483	23.209	25.188	27.112	29.588	31.420
11	14.631	17.275	19.675	21.920	24.725	26.757	28.729	31.264	33.136
12	15.812	18.549	21.026	23.337	26.217	28.300	30.318	32.909	34.821
13	16.985	19.812	22.362	24.736	27.688	29.819	31.883	34.528	36.478
14	18.151	21.064	23.685	26.119	29.141	31.319	33.426	36.123	38.109
15	19.311	22.307	24.996	27.488	30.578	32.801	34.950	37.697	39.719
16	20.465	23.542	26.296	28.845	32.000	34.267	36.456	39.252	41.308
17	21.615	24.769	27.587	30.191	33.409	35.718	37.946	40.790	42.879
18	22.760	25.989	28.869	31.526	34.805	37.156	39.422	42.312	44.434
19	23.900	27.204	30.144	32.852	36.191	38.582	40.885	43.820	45.974
20	25.028	28.412	31.410	34.170	37.566	39.997	42.336	45.315	47.498
21	26.171	29.615	32.671	35.479	38.932	41.401	43.775	46.797	49.011
22	27.301	30.813	33.924	36.781	40.289	42.796	45.204	48.268	50.511
23	28.429	32.007	35.172	38.076	41.638	44.181	46.623	49.728	52.000
24	29.553	33.196	36.415	39.364	42.980	45.558	48.034	51.179	53.478
25	30.675	34.382	37.652	40.646	44.314	46.928	49.435	52.620	54.947
26	31.795	35.563	38.885	41.923	45.642	48.290	50.829	54.052	56.407
27	32.912	36.741	40.113	43.195	46.963	49.645	52.215	55.476	57.857
28	34.027	37.916	41.337	44.461	48.278	50.993	53.594	56.892	59.300
29	35.139	39.087	42.557	45.722	49.588	52.336	54.967	58.301	60.734
30	36.250	40.256	43.773	46.979	50.892	53.672	56.332	59.703	62.162
31	37.359	44.422	44.985	48.232	52.191	55.003	57.692	61.098	63.582
32	38.466	42.585	46.194	49.480	53.486	56.328	59.046	62.487	64.995
33	39.572	43.745	47.400	50.725	54.776	57.648	60.395	63.870	66.402
34	40.676	44.903	48.602	51.966	56.061	58.964	61.738	65.247	67.803
35	41.779	46.059	49.802	53.203	57.342	60.275	63.076	66.619	69.198
36	42.879	47.212	50.998	54.437	58.619	61.581	64.410	67.985	70.588
37	43.978	48.363	52.192	55.668	59.892	62.883	65.739	69.346	71.972
38	45.076	49.513	53.384	56.896	61.162	64.181	67.063	70.703	73.351
39	46.173	50.660	54.572	58.120	62.428	65.476	68.383	72.055	74.725
40	47.269	51.805	55.758	59.342	63.691	66.766	69.699	73.402	76.094
45	52.729	57.505	61.656	65.410	69.957	73.166	76.233	80.077	82.875
50	58.164	63.167	67.505	71.420	76.154	79.490	82.664	86.661	89.560
60	68.970	74.399	79.087	83.305	88.386	91.957	95.357	99.607	102.689
70	79.712	85.529	90.537	95.031	100.432	104.222	107.812	112.319	115.575
80	90.403	96.581	101.885	106.636	112.336	116.329	120.107	124.842	128.261
90	101.051	107.568	113.151	118.144	124.125	128.307	132.262	137.213	140.783
100	111.664	118.501	124.348	129.570	135.815	140.178	144.300	149.455	153.169
110	122.247	129.388	135.487	140.925	147.423	151.958	156.238	161.587	165.439
120	140.231	146.571	152.222	157.389	163.678	168.122	172.351	177.673	181.528
140	161.826	168.618	174.659	180.174	186.875	191.604	196.099	201.748	205.835
160	183.310	190.522	196.926	202.766	209.852	214.845	219.588	225.542	229.846
180	204.704	212.310	219.056	225.200	232.647	237.890	242.866	249.107	253.615

Appendix 4

Random numbers

1735	6040	2537	5480	9607	7165	8376	7704	6253	8711	6338	0933	3734	3541	8013
3261	8742	2304	9303	7416	0565	5450	4154	6596	8879	6744	0285	0510	8070	3515
9259	5782	4890	8924	1708	8867	1952	1557	4592	8362	4715	3392	4152	1515	1212
4035	7559	8763	7540	8831	4679	2634	9421	4160	7124	3779	4261	4552	2777	2567
0290	4533	3135	6361	9181	8035	8864	8848	1910	6995	9393	3668	6865	0907	2540
1164	4842	2873	6089	9329	7601	5677	7791	5219	7374	6237	5750	0175	5226	9720
5966	3457	8758	0895	4598	8470	4230	6950	9633	5212	6010	3953	5994	7137	1089
3141	9842	8447	7162	3588	0899	1051	1157	7245	1020	0524	6272	9182	8761	3740
1252	8064	3481	4190	1143	6387	7079	2801	0159	1781	0733	7198	8739	7092	3640
6978	4272	1341	7000	7980	2319	2584	5282	5958	1674	4146	3629	7730	9532	5685
1299	0796	7496	7440	4156	6879	4664	2674	0835	5061	7999	2398	7383	5947	3686
9926	0374	0643	8959	8106	3343	8217	6471	2277	4697	1634	2177	8672	2312	5497
3712	7751	4376	7986	1891	8062	1276	7815	0532	1335	7942	1965	0922	8934	7233
2762	5147	0411	1731	3913	8593	7340	0314	9319	2465	0271	4302	6616	5774	2501
8905	2781	3558	6024	6778	6340	3366	0465	9142	4588	4658	2185	3827	5733	9626
7389	6272	8985	9127	4010	6312	3424	4285	1721	7982	4645	6455	8196	1428	7362
9846	6925	9103	1047	6084	4003	9758	9522	2662	0821	9328	8993	5434	4996	5331
0352	6475	9070	4029	6023	3599	3007	8120	0180	8357	8349	3565	8454	6430	8826
7913	5974	9943	9689	9300	3874	3858	7304	5401	2088	9099	9628	3620	3469	6848
8351	9866	8042	6620	7985	5611	3716	4181	8707	8536	6489	4453	8728	2647	6783
2443	9757	3987	0509	8441	7147	8163	4252	2191	6920	6796	2642	2022	2540	3618
3255	7382	7078	8600	6781	4543	6331	4214	4213	5701	8048	7996	9583	4771	5976
2920	7022	5141	0821	9634	4175	5380	5691	3842	4360	2912	8560	8947	8765	6318
3654	9193	4711	3553	9797	7351	6750	3395	5892	4753	1851	1229	0184	1788	3843
5767	3354	3308	0792	0753	3594	0643	8561	8546	3808	4059	8198	7335	2333	1988
7796	2040	1922	3943	1375	0716	0426	5486	8943	1856	3922	5899	6190	9420	0560
7342	5651	9066	4897	6809	5340	8932	0719	9260	3084	9338	3583	5209	2690	9763
3361	6102	7408	6675	0037	0524	5463	1705	0931	0663	7990	8546	4899	2869	6268
3506	2001	6497	0880	2569	3728	3759	0292	2291	9912	5016	7780	7499	1987	9732
6525	7563	2468	9127	3407	8261	5075	6392	7974	1029	8040	6870	4390	7812	7181
1535	4491	9896	2736	5931	7094	3650	2935	0643	8813	6896	0774	3275	6583	9742
1443	5427	3403	1525	7027	9445	0859	8626	2717	5805	3989	7985	6057	8630	8888
3179	9771	9654	1384	6747	9815	7174	3310	5041	5453	7162	2114	8826	9008	2872
1837	2212	1857	0660	4132	1851	1264	9426	0338	6420	3574	1714	7933	9386	5282
5290	9901	0274	5198	0554	9806	1503	2387	7805	7553	9313	4437	1244	7682	1882
0212	9446	3011	2592	1310	4465	7825	9503	3931	3462	3261	9340	8012	1277	8401
7091	4956	2054	2691	8882	2907	6151	7517	3280	9513	4702	2844	4825	5581	9386
4677	1224	3109	8451	3782	4842	2467	5606	8009	9172	1462	5289	9855	5378	8857
5485	8074	4291	1514	0791	3314	1499	9650	8902	8800	5023	9381	0039	5692	1374
4900	0071	4292	3506	1749	1638	8503	7873	0659	1823	5811	6169	5656	7311	7140
8319	9210	2062	5618	6247	9514	0616	5893	5534	1776	4014	4867	7869	8465	9638
9565	2204	3553	7383	5048	7917	0087	2663	2293	3705	8220	2796	2441	3192	8671
1473	7793	8439	7219	4046	5971	0864	7520	5644	7943	1333	8080	5665	5908	0022
9729	3259	3822	1355	2759	5663	7467	0628	2813	4834	7558	6209	7529	5300	4290
3546	9220	9008	2460	3418	5320	7195	3316	3489	8999	7847	1261	0844	4529	9174

Statistical Methods and Calculation Skills

Third edition

This third edition of *Statistical Methods and Calculation Skills* aims to equip students with the skills to apply statistical analysis and quantitative techniques to research and in the working environment and to make effective decisions.

The book provides:

- A theoretical framework for statistical problem-solving
- A practical step-by-step approach to applying methods and calculations
- A complete list of outcomes in each unit
- Worked examples with detailed explanations
- Practice in the form of guided activities and a range of self-test questions.

The contents include the collection and presentation of data, descriptive measures, index numbers, regression and correlation analysis, time series, probability and probability distributions, statistical estimation and hypothesis testing. Calculation skills are revised in Part 2, a section that covers technology, elementary calculations, percentages and ratios, equations, graph construction and interest calculations.

This edition includes examples and activities from the fields of business, food and biotechnology, engineering, medicine and environmental studies.

www.juta.co.za



JUTA

Copyrighted material