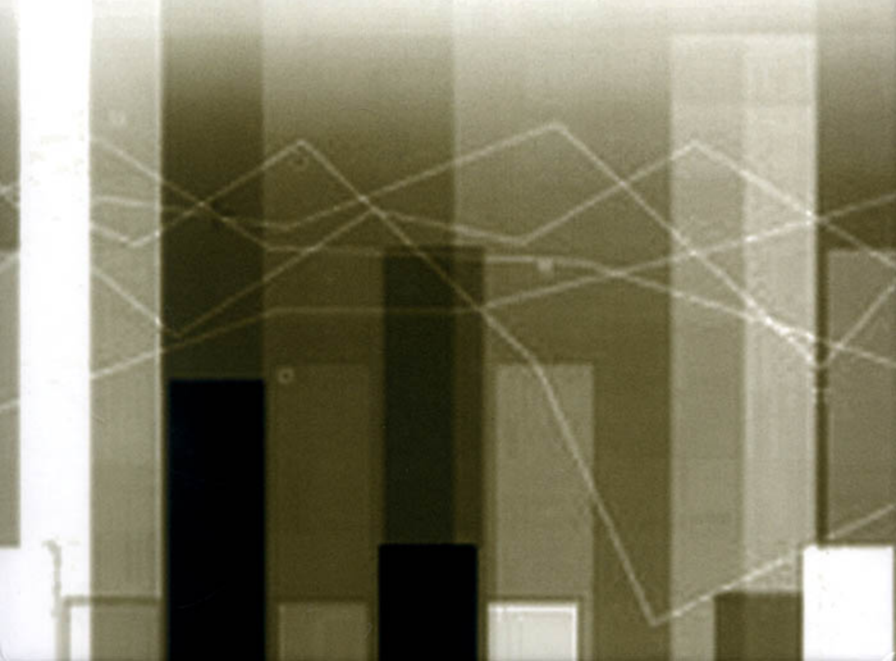
THE NEW SPACE ERA

SECOND EDITION

STATISTICAL POWER ANALYSIS

A Simple and General Model for Traditional and Modern Hypothesis Tests

KEVIN R. MURPHY
BRETT MYORS



Statistical Power Analysis

A Simple and General Model
for Traditional and Modern Hypothesis Tests

Second Edition

This page intentionally left blank

Statistical Power Analysis

A Simple and General Model for Traditional and Modern Hypothesis Tests

Second Edition

Kevin R. Murphy
Pennsylvania State University

Brett Myors
Griffith University

2004 LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
Mahwah, New Jersey London

Lawrence Erlbaum Associates, Inc., Publishers
10 Industrial Avenue
Mahwah, NJ 07430

Cover design by Kathryn Houghtaling Lacey

Library of Congress Cataloging-in-Publication Data

Statistical power analysis : a simple and general model for traditional and modern hypothesis tests, second edition, by Kevin R. Murphy and Brett Myors.

Includes bibliographical references and index.
ISBN 0-8058-4525-9 (cloth : alk. paper)
ISBN 0-8058-4526-7 (pbk. : alk. paper)

Copyright information for this volume can be obtained by contacting the Library of Congress

2004
—dc21 2004000000
CIP

Books published by Lawrence Erlbaum Associates are printed on acid-free paper, and their bindings are chosen for strength and durability.

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

Disclaimer:
This eBook does not include the ancillary media that was packaged with the original printed version of the book.

Contents

Preface

One of the most common statistical procedures in the behavioral and social sciences is to test the hypothesis that treatments or interventions have no effect, or that the correlation between two variables is equal to zero, and so on (i.e., tests of the null hypothesis). Researchers have long been concerned with the possibility that they will reject the null hypothesis when it is in fact correct (i.e., make a Type I error), and an extensive body of research and data-analytic methods exists to help understand and control these errors. Substantially less attention has been devoted to the possibility that researchers will fail to reject the null hypothesis, when in fact treatments, interventions, and so forth. have some real effect (i.e., make a Type II error). Statistical tests that fail to detect the real effects of treatments or interventions might substantially impede the progress of scientific research.

The statistical power of a test is the probability that it will lead you to reject the null hypothesis when that hypothesis is in fact wrong. Because most statistical tests are done in contexts where treatments have at least some effect (although it might be minuscule), power often translates into the probability that the test will lead to a correct conclusion about the null hypothesis. Viewed in this light, it is obvious why researchers have become interested in the topic of statistical power, and in methods of assessing and increasing the power of their tests.

This book presents a simple and general model for statistical power analysis based on the widely used *F* statistic. A wide variety of statistics used in the social and behavioral sciences can be thought of as special applications of the general linear model (e.g., *t* tests, analysis of variance and covariance, correlation, multiple regression), and the *F* statistic can be used in testing hypotheses about virtually any of these specialized applications. The model for power analysis laid out here is quite simple, and it illustrates how these analyses work and how they can be applied to problems of study design, to evaluating others' research, and even to problems such as choosing the appropriate criterion for defining statistically significant outcomes.

In response to criticisms of traditional null hypothesis testing, several researchers have developed methods for testing what is referred to as a minimum-effect hypothesis (i.e., the hypothesis that the effect of treatments, interventions, etc. exceeds some specific minimal level). This is the first book to discuss in detail the application of power analysis to both traditional null hypothesis tests and to minimum-effect tests. It shows how the same basic model applies to both types of testing, and illustrates applications of power analysis to both traditional null hypothesis tests (i.e., tests of the hypothesis that treatments have no effect) and to minimum-effect tests (i.e., tests of the hypothesis that the effects of treatments exceeds some minimal level). A single table is used to conduct both significance tests and power analyses for traditional and for minimum-effect tests (The One-Stop *F* Table, presented in Appendix B), and some relatively simple procedures are presented that may be used to ask a series of important and sophisticated questions about the research.

This book is intended for a wide audience, and so presentations are kept simple and nontechnical wherever possible. For example, Appendix A presents some fairly daunting statistical formulas, but it also shows how a researcher with little expertise or interest in statistical analysis could quickly obtain the values needed to carry out power analyses for any range of hypotheses. Similarly, the first three chapters of this book present a few formulas, but the reader who skips them entirely will still be able to follow the ideas being presented in this book.

Finally, most of the examples presented herein are drawn from the social and behavioral sciences, as are many of the generalizations about statistical methods that are most likely to be used. In part, this reflects our biases (we are both psychologists), but it also reflects the fact that issues related to power analysis have been widely discussed in this literature over the last several years. Researchers in other areas may find that some of the specific advice offered here does not apply as well to them, but the general principles articulated in this book should be useful to researchers in a wide range of disciplines.

This second edition includes a number of features that were not part of our first edition. First, a chapter (chap. 4) dealing with power analysis in multifactor analysis of variance (ANOVA), including repeated measures designs, has been added. Multifactor ANOVA is very common in the behavioral and social sciences, and whereas the conceptual issues in power analysis are quite similar in factorial ANOVA as in other methods of analysis, there are several features of ANOVA that require special attention, and this topic deserves treatment in a separate chapter. Second, a "One-Stop *PV* Table" has been included, which presents the same information as in the One-Stop *F* Table, framed in terms of the percentage of variance (*PV*) explained rather than in terms of *F*. This table allows researchers to find a quick and simple answer to questions like "How large would the effect have to be in my study to yield power of .80?". Finally, a CD with a simple program called the "One-Stop *F* Calculator" is included, which allows researchers to find both *F* and *PV* values needed for testing hypotheses and for estimating power for both traditional and modern hypothesis testing strategies. This program makes it possible to put the concepts in this book into play quickly and easily.

This page intentionally left blank

1

The Power of Statistical Tests

In the social and behavioral sciences, statistics serve two general purposes. First, they can be used to describe what happened in a particular study (descriptive statistics). Second, they can be used to help draw conclusions about what those results mean in some broader context (inferential statistics). The main question in inferential statistics is whether a result, finding, or observation from a study reflects some meaningful phenomenon in the population from which that study was drawn. For example, if 100 college sophomores are surveyed and it is determined that a majority of them prefer pizza to hot dogs, then does this mean that people in general (or college students in general) also prefer pizza? If a medical treatment yields improvements in 6 out of 10 patients, then does this mean that it is an effective treatment that should be approved for general use?

The process of drawing inferences about populations from samples is a risky one, and a great deal has been written about the causes and cures for errors in statistical inference. Statistical power analysis (J. Cohen, 1988; Kraemer & Thiemann, 1987; Lipsey, 1990) falls under this general heading. Studies with too little statistical power can frequently lead to erroneous conclusions. In particular, they will very often lead to the incorrect conclusion that findings reported in a particular study are not likely to be true in the broader population. In the previous example, the fact that a medical treatment worked for 6 out of 10 patients is probably insufficient evidence that it is truly safe and effective, and if there is nothing more than this study to rely on, it might be concluded that the treatment has not been proven effective.

This conclusion may say as much about the low level of statistical power in the study as about the value of the treatment.

This chapter describes the rationale for and applications of statistical power analysis. In most of examples, it describes or applies power analysis in studies that assess the effect of some treatment or intervention (e.g., psychotherapy, reading instruction, performance incentives) by comparing outcomes for those who have received the treatment to outcomes of those who have not (nontreatment or control group). However, as is emphasized throughout this book, power analysis is applicable to a very wide range of statistical tests, and the same simple and general model can be applied to many of the statistical techniques used in the social and behavioral sciences.

THE STRUCTURE OF STATISTICAL TESTS

Understanding statistical power requires first understanding the ideas that underlie statistical hypothesis testing. Suppose 50 children are exposed to a new method of reading instruction, and then it is shown that their performance on reading tests is, on average, 6 points higher (on a 100-point test) than that of 50 similar children who received standard methods of instruction. Does this mean that the new method is truly better? A 6-point difference might mean that the new method is really better, but it is also possible that there is no real difference between the two methods, and this observed difference is the result of the sort of random fluctuation that might be expected when the results from a single sample are used (here, the 100 children assigned to the two reading programs) to draw inferences about the effects of these two methods of instruction in the population.

One of the most basic ideas in statistical analysis is that results obtained in a sample do not necessarily reflect the state of affairs in the population from which that sample was drawn. For example, the fact that scores averaged 6 points higher in this particular group of children does not necessarily mean that scores will be 6 points higher in the population, or that the same 6-point difference would be found in another study examining a new group of students. Because samples do not (in general) perfectly represent the populations from which they were drawn, some instability should be expected in the results obtained from each sample. This instability is usually referred to as *sampling error.* The presence of sampling error is what makes drawing inferences about populations from samples risky. One of the key goals of statistical theory is to estimate the amount of sampling error likely to be present in different statistical procedures and tests, and thereby gaining some idea about the amount of risk involved in using a particular procedure.

Statistical significance tests can be thought of as decision aids. That is, these tests can help researchers draw conclusions about whether the findings of a particular study represent real population effects, or whether they fall within the range of outcomes that might be produced by random sampling error. For example, there are two possible interpretations of the findings in this study of reading instruction:

1. The difference between average scores from the two programs is so small that it might reasonably represent nothing more than sampling error.
2. The difference between average scores from the two programs is so large that it cannot be reasonably explained in terms of sampling error.

The most common statistical procedure in the social and behavioral sciences is to pit a null hypothesis (H_0) against an alternative (H_1). In this example, the null and alternative hypotheses might take the following forms:

H_0—Reading instruction has no effect. It does not matter how children are taught to read, because in the population there is no difference in the average scores of children receiving either method of instruction.

H_1—Reading instruction has an effect. It does matter how you teach children are taught to read, because in the population there is a difference in the average scores of children receiving different method of instruction.

Although null hypotheses usually refer to "no difference" or "no effect," it is important to understand that there is nothing magic about the hypothesis that the difference between two groups is zero. It might be perfectly reasonable to evaluate the following set of possibilities:

H_0—In the population, the difference in the average scores of those receiving these two methods of reading instruction is 6 points.

H_1—In the population, the difference in the average scores of those receiving these two methods of reading instruction is not 6 points.

The *null hypothesis* (H_0) is a specific statement about results in a population that can be tested (and therefore nullified). One reason that null hypotheses are often framed in terms of "no effect" is that the alternative that is implied by this hypothesis is easy to interpret.

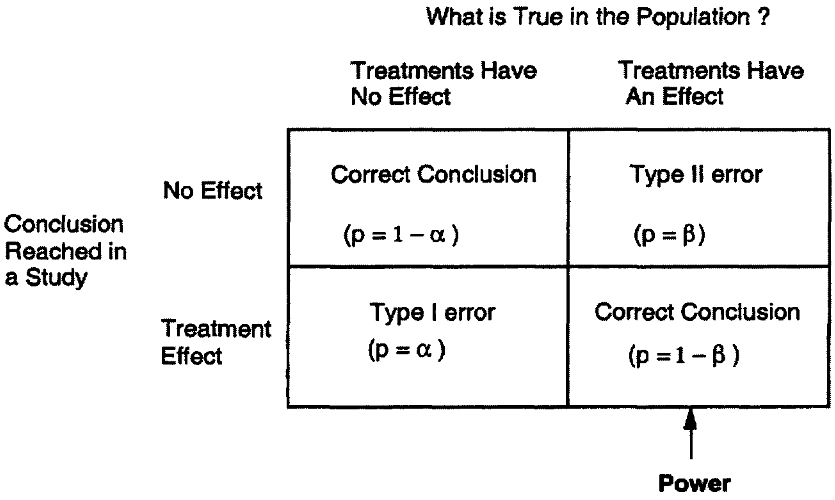
If researchers test and reject the hypothesis that treatments have no effect, they are left with the alternative that treatments have at least some effect. Another reason for testing the hypothesis that treatments have no effect whatsoever is that probabilities, test statistics, and so on, are easy to calculate when the effect of treatments is assumed to be zero.

In contrast, if researchers test and reject the hypothesis that the difference between treatments is exactly 6 points, they are left with a wide range of alternatives (e.g., the difference is 5 points, the difference is 10 points, etc.), including the possibility that there is no difference whatsoever. Although the hypothesis that treatments have no effect is the most common basis for statistical hypothesis tests (J. Cohen, 1994, refers to this hypothesis as the "nil hypothesis"), as is shown later, there are a number of advantages to posing and testing substantive hypotheses about the size of treatment effects (Murphy & Myors, 1999). For example, it is easy to test the hypothesis that the effects of treatments are negligibly small (e.g., they account for 1% or less of the variance in outcomes, or that the standardized mean difference is .10 or less). If researchers test and reject this hypothesis, they are left with the alternative hypothesis that the effect of treatments is not negligibly small, but rather is large enough to deserve at least some attention. The methods of power analysis described in this book are easily extended to such minimum-effect tests, and are not in any way limited to traditional tests of the null hypothesis that treatments have no effect.

What Determines the Outcomes of Statistical Tests?

There are four outcomes that can occur when researchers use the results obtained in a particular sample (e.g., the finding that one treatment works better than another in that sample) to draw inferences about a population (e.g., the inference that the treatment will also be better in the population). These outcomes are shown in Fig. 1.1.

The concern here is with understanding and minimizing errors in statistical inference; as Fig. 1.1 shows, there are two ways to make errors when testing hypotheses. First, it is possible that the treatment (e.g., new method of instruction) has no real effect in the population, but the results in the sample might lead to the belief that it does have some effect. If you the results of this study were used to conclude that the new method of instruction was truly superior to the standard method, when in fact there were no differences, then this would be a *Type I error*. Type I errors might lead researchers to waste time and resources by pursuing what is essentially a dead end, and researchers have traditionally gone to great lengths to avoid these errors.



		What is True in the Population ?	
		Treatments Have No Effect	Treatments Have An Effect
Conclusion Reached in a Study	No Effect	Correct Conclusion $(p = 1 - \alpha)$	Type II error $(p = \beta)$
	Treatment Effect	Type I error $(p = \alpha)$	Correct Conclusion $(p = 1 - \beta)$

FIG. 1.1 Outcomes of statistical tests.

There is an extensive literature dealing with methods of estimating and minimizing the occurrence of Type I errors (Zwick & Marascuilo, 1984). The probability of making a Type I error is in part a function of the standard or decision criterion used in testing a hypothesis (often referred to as alpha, or a). A very lenient standard (e.g., if there is any difference between the two samples, it will be concluded that there is also a difference in the population) might lead to more frequent Type I errors, whereas a more stringent standard might lead to few Type I errors.[1]

A second type of error, referred to as Type II error, is also common in statistical hypothesis testing (J. Cohen, 1994; Sedlmeier & Gigerenzer, 1989). A Type II error occurs when researchers conclude in favor of H_0, when in fact H_1 is true. Statistical power analysis is concerned with Type II errors. The power of a statistical test is defined as one minus the probability of making a Type II error (i.e., if the probability of making a Type II error is β, power = 1 − β, or power is the probability that you will avoid a Type II error). Studies with high levels of statistical power will rarely fail to detect the effects of treatments. If it

[1] It is important to note that Type I errors can only occur when the null hypothesis is actually true. If the null hypothesis is that there is no true treatment effect (a nil hypothesis), then this will rarely be the case. As a result, Type I errors are probably quite rate in tests of the traditional null hypothesis, and efforts to control these errors at the expense of making more Type II errors might be ill advised (Murphy, 1990).

is assumed that most treatments have at least some effect, then the statistical power of a study translates into the probability that the study will lead to the correct conclusion (i.e, that it will detect the effects of treatments).

Effects of Sensitivity, Effect Size, and Decision Criteria on Power

The power of a statistical test is a function of its sensitivity, the size of the effect in the population, and the standards or criteria used to test statistical hypotheses. Studies have higher levels of statistical power when:

1. They are highly sensitive. Researchers may increase sensitivity by using better measures, or a study design that allows them to control for unwanted sources of variability in the data (for the moment, sensitivity is defined in terms of the degree to which sampling error introduces imprecision into the results of a study; a fuller definition is presented later in this chapter). The simplest method of increasing the sensitivity of a study is to increase its sample size (N). As N increases, statistical estimates become more precise and the power of statistical tests increase.

2. Effect sizes (*ES*) are large. Different treatments have different effects. It is easiest to detect the effect of a treatment if that effect is large (e.g., when treatment means are very different, or, relatedly, when treatments account for a substantial proportion of variance in outcomes; specific measures of effect size are discussed later in this chapter and in the chapters that follow). When treatments have very small effects, these effects can be difficult to reliably detect. As *ES* values increase, power increases.

3. Standards are set that make it easier to reject H_0. It is easier to reject H_0 when the significance criterion, or alpha (α) level, is .05 than when it is .01 or .001. As the standard for determining significance becomes more lenient, power increases.

Power is highest when all three of these conditions are met (i.e., sensitive study, large effect, lenient criterion for rejecting the null hypothesis). In practice, sample size (which affects sensitivity) is probably the more important determinant of power. Effect sizes in the social and behavioral sciences tend to be small or moderate (if the effect of a treatment is so large that it can be seen by the naked eye, even in small samples, then there may be little reason to test for it statistically), and researchers are often unwilling to abandon the traditional criteria for statistical significance that are accepted in their field (usually, alpha levels of .05 or .01; Cowles & Davis, 1982). Thus, effect sizes and decision criteria tend to be similar across a wide range of studies. In contrast, sample sizes vary considerably, and they directly impact levels of power. With a sufficiently large *N*, virtually any statistic will be "significantly" different from zero, and virtually any null hypothesis that is tested will be rejected. Large *N* makes statistical tests highly sensitive, and virtually any specific point hypothesis can be rejected if the study is sufficiently sensitive (as is shown later, this is not true for tests of the hypothesis that treatment effects fall in some range of values defined as "negligibly small" or "meaningfully large"). For example, if the effect of a new medication is an increase of .000001% in the success rate of treatments, then the null hypothesis that treatments have no effect is formally wrong, and will be rejected in a study that is sufficiently sensitive. With a small enough *N*, there may not be enough power to reliably detect the effects of even the most substantial treatments.

Studies can have very low levels of power (i.e., are likely to make Type II errors) when they use small samples, when the effect being studied is a small one, and/or when stringent criteria are used to define a significant result. The worst case occurs when researchers use a small sample to study a treatment that has a very small effect, and they use a very strict standard for rejecting the null hypothesis. Under those conditions, Type II errors may be the norm. To put it simply, studies that use small samples and stringent criteria for statistical significance to examine treatments that have small effects will almost always lead to the wrong conclusion about those treatments (i.e., to the conclusion that treatments have no effect whatsoever).

THE MECHANICS OF POWER ANALYSIS

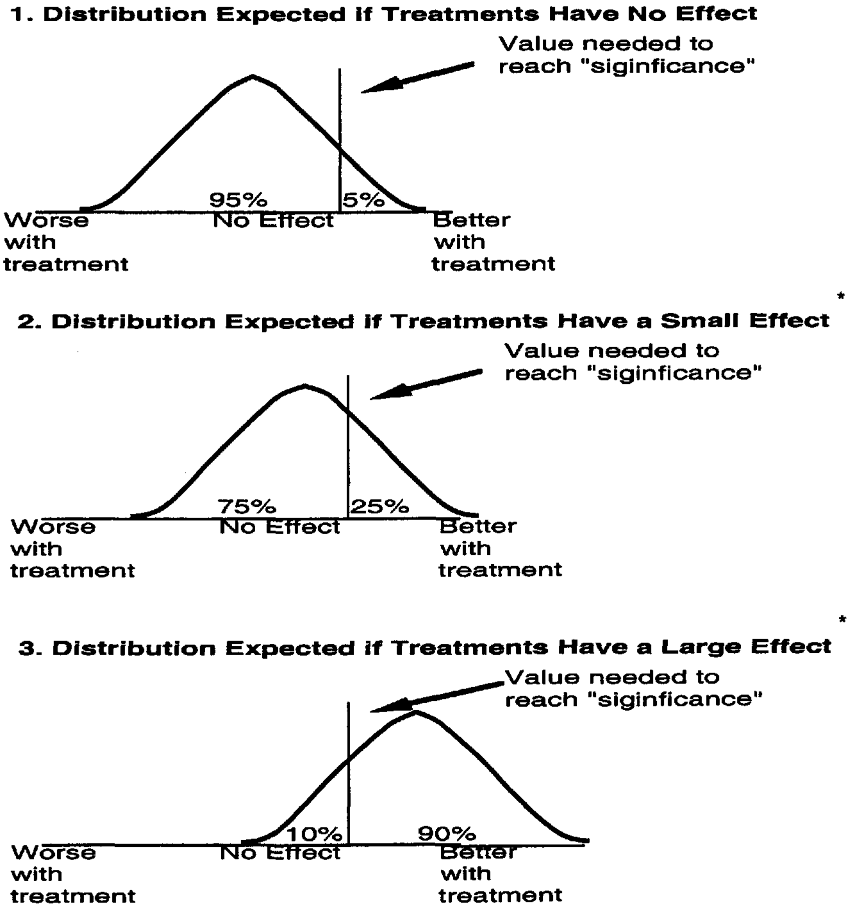
When a sample is drawn from a population, the exact value of any statistic (e.g., the mean, difference between two group means) is uncertain, and that uncertainty is reflected by a statistical distribution. Suppose, for example, that researchers introduce a treatment that has no real effect (e.g., they use astrology to advise people about career choices), and then compare outcomes for groups who receive this treatment to outcomes for groups who do not receive it (control groups). They will not always find that treatment and control groups have exactly the same scores, even if the treatment has no real effect. Rather, there is some range of values they might expect for any test statistic in a study like this, and the standards used to determine statistical significance are based on this range or distribution of values. In traditional null hypothesis testing, a test statistic is statistically significant at the .05 level if its actual value is outside of the range of values they would observe 95% of the time in studies where the treatment had no real effect. If the test statistic is outside of this range, then the inference is that the treatment did have some real effect.

For example, suppose that 62 people are randomly assigned to treatment and control groups, and the *t* statistic is used to compare the means of the two groups. If the treatment has no effect whatsoever, the *t* statistic should usually be near zero, and will have a value less than or equal to 2.00 in 95% of all such studies. If the *t* statistic obtained in a study is larger than 2.00, then the inference is that treatments are very likely to have some effect; if there was no real effect of treatments, then values above 2.00 would be a very rare event.

As the pervious example suggests, if treatments have no effect whatsoever in the population, researchers should not expect to always find a difference of precisely zero between samples of those who do and do not receive the treatment. Rather, there is some range of values that might be found for any test statistic in a sample (e.g., in the example cited earlier, the value of a *t* statistic is expected to be near zero, but it might range from –2.00 to +2.00). The same is true if treatments have a real effect. For example, if researchers expect that the mean in a treatment group will be 10 points higher than the mean in a control group (e.g., because this is the size of the difference in the population), they should also expect some variability around that figure. Sometimes, the difference between two samples might be 9 points, and sometimes it might be 11 or 12 points. The key to power analysis is estimating the range of values to reasonably expect for some test statistic if the real effect of treatments is small, or medium, or large.

Figure 1.2 illustrates the key ideas in statistical power analysis. Suppose researchers devise a new test statistic and use it to evaluate the 6-point difference in reading test scores described earlier. The larger the difference between the two treatment groups, the larger the value of their test statistic. To be statistically significant, the value of this test statistic must be 2.00 or larger. As Fig. 1.2 suggests, the chance they will reject the null hypothesis that there is no difference between the two groups depends substantially on whether the true effect of treatments is small or large.

If the null hypothesis that there is no real effect was true, then they would expect to find values of 2.00 or higher for this test statistic in 5 tests out of every 100 performed (i.e., $\alpha = .05$). This is illustrated in Section 1 of Fig. 1.2. Section 2 of Fig. 1.2 illustrates the distribution of test statistic values they might expect if treatments had a small effect on the dependent variable. Researchers might notice that the distribution of test statistics they would expect to find in studies of a treatment with this sort of effect has shifted a bit, and in this case 25% of the values they might expect to find are greater than or equal to 2.00. That is, if the study is run under the scenario illustrated in Section 2 of this fig-



1. Distribution Expected if Treatments Have No Effect

2. Distribution Expected if Treatments Have a Small Effect *

3. Distribution Expected if Treatments Have a Large Effect *

* - depending on the test statistic in question, the distributions might take different forms, but the essential features of this figure would apply to any test statistic

FIG. 1.2. Essentials of power analysis.

ure (i.e., treatments have a small effect), then the probability researchers will reject the null hypothesis is .25. Section 3 of Fig. 1.2 illustrates the distribution of values they might expect if the true effect of treatments is large. In this distribution, 90% of the values are 2.00 or greater, and the probability they will reject the null hypothesis is .90. The power of a statistical test is the proportion of the distribution of test statistics expected for that study that is above the critical value used to establish statistical significance.

No matter what hypothesis is being tested or what statistic researchers are using to test that hypothesis, power analysis always involves three basic steps, listed in Table 1.1. First, you must set some criterion or critical value for "statistical significance". For example, the tables found in the back of virtually any statistics textbook can be used to determine such critical values for testing the traditional null hypothesis. If the test statistic computed exceeds this critical value, you will reject the null. However, these tables are not the only basis for setting such a criterion. Suppose you want to test the hypothesis that the effects of treatments are so small that they can safely be ignored. This might involve specifying

TABLE 1.1

The Three Steps to Determining Statistical Power

1. Establish a criterion or critical value for statistical significance:
   - What is the hypothesis (e.g., traditional null hypothesis,, minimum-effect tests)?
   - What level of confidence is desired (e.g., $\alpha = .05$ vs. $\alpha = .01$)?
   - What is the critical value for the test statistic (based on the degrees of freedom for the test and the " level)?
2. Estimate the effect size:
   - Do you expect the treatments to have a large,, medium,, or small effect?
   - What is the range of values researchers expect to find for their test statistic, given this effect size?
3. Determine where the critical value lies in relationship to the distribution of test statistics researchers expect to find in a study:
   - The power of a statistical test is the proportion of the distribution of test statistics expected for that study that is above the critical value used to establish statistical significance.

some range of effects that would be designated as "negligible," and then determining the critical value of a statistic needed to reject this hypothesis. Chapter 2 shows the way such tests are done, and the implications of such hypothesis testing strategies for statistical power analysis.

Second, an effect size must be estimated. That is, researchers must make their best guess of how much effect the treatments being studied are likely to have on the dependent variable(s); methods of estimating effect sizes are discussed later in this chapter. As noted earlier, if there are good reasons to believe that treatments have a very large effect, then it should be quite easy to reject the null hypothesis. On the other hand, if the true effects of treatments are small and subtle, then it might be very hard to reject the hypothesis that they have no real effect.

Once researchers have estimated the effect size, it is also possible to use that estimate to describe the distribution of test statistics they should expect to find in studies of that particular treatment or set of treatments. This process is described in more detail in chapter 2, but a simple example serves to illustrate. Suppose researchers are using the *t* test to assess the difference in the mean scores of those receiving two different treatments. If there was no real difference between the treatments, then they would expect to find *t* values near zero most of the time, and they could use statistical theory to tell how much these values might depart from zero as a result of sampling error. The *t* tables in most statistics textbooks tell how much variability researchers might expect with samples of different sizes, and once they know the mean (here, zero) and the standard deviation of this distribution, it is easy to estimate what proportion of the distribution falls above or below any critical value. If there is a large difference between the treatments (e.g., the dependent variable has a mean of 500 and a standard deviation of 100, and the mean for one treatment is usually 80 points higher than the mean for another), then they should expect to find large *t* values most of the time, and once again, they can use statistical theory to estimate the distribution of values expected in such studies.

The final step in power analysis is a comparison between the values obtained in the first two steps. For example, if it is determined that a *t* value of 2.00 is needed to reject a particular null hypothesis, and it is also determined that because the treatments being studied have very large effects it is likely that *t* values of 2.00 or greater will be found 90% of the time, then the power of this test (i.e., power is .90) has also been determined.

Sensitivity and Power

Sensitivity refers to the precision with which a statistical test distinguishes between true treatment effects and differences in scores that are the result of sampling error. As already noted, the sensitivity of statistical tests is largely a function of the sample size. Large samples provide very precise estimates of population parameters, whereas small samples produce results than can be unstable and untrustworthy. For example, if 6 children in 10 do better with a new reading curriculum than with the old one, this might reflect nothing more than simple sampling error. If 600 of 1,000 children do better with the new curriculum, this is powerful and convincing evidence that there are real differences between the new curriculum and the old one. In a study with low sensitivity, there is considerable uncertainty about statistical outcomes. As a result, it might be possible to find a large treatment effect in a sample, even though there is no true treatment effect in the population. This translates into substantial variability in study outcomes and the need for relatively demanding tests of statistical significance. If outcomes can vary substantially from study to study, researchers need to observe a relatively large effect to be confident that it represents a true treatment effect and not merely sampling error. As a result, it will be difficult to reject the hypothesis that there is no true effect, and many Type II errors might be made.

In a highly sensitive study, there is very little uncertainty or random variation in study outcomes, and virtually any difference between treatment and control groups is likely to be accepted as an indication that the treatment has an effect in the population.

Effect Size and Power

Effect size is a key concept in statistical power analysis (J. Cohen, 1988; Rosenthal, 1991; Tatsuoka, 1993a). At the simplest level, effect size measures provide a standardized index of how much impact treatments actually have on the dependent variable. One of the most common effect size measures is the standardized mean difference, *d*, defined as $d = (M_t - M_c)/SD$, where M_t and M_c are the treatment and control group means, respectively, and *SD* is the pooled standard deviation. By expressing the difference in group means in standard deviation units, the *d* statistic provides a simple metric that allows for comparison of treatment effects from different studies, areas or research, and so on, without having to keep track of the units of measurement used in different studies or areas of research. For example, Lipsey and Wilson (1993) cataloged the effects of a wide range of psychological, educational, and behavioral treatments, all expressed in terms of *d*. Examples of interventions in these areas that have relatively small, moderately large, and large effects on specific sets of outcomes are presented in Table 1.2.

For example, worksite smoking cessation/reduction programs have a relatively small effect on quit rates ($d = .21$). The effects of class size on achievement or of juvenile delinquency programs on delinquency outcomes are similarly small. Concretely, a d value of .20 means that the difference between the average score of those who receive the treatment and those who do not is only 20% as large as the standard deviation of the outcome measure within each of the treatment groups. This standard deviation measures the variability in outcomes, independent of treatments, so $d = .20$ indicates that the average effect of treatments is only one fifth as large as the variability in outcomes that might be seen with no treatments. In contrast, interventions such as psychotherapy, meditation and relaxation, or positive reinforcement in the classroom have relatively large effects on

TABLE 1.2

Examples of Effect Sizes Reported in Lipsey and Wilson (1993) Review

Effect Size	*Dependent Variable*	*d*
Small (d = .20)		
Treatment programs for juvenile delinquents	delinquency outcomes	.17
Worksite smoking cessation/reduction programs	quit rates	.21
Small vs. large class size, all grade levels	achievement measures	.20
Medium (d = .50)		
Behavior therapy vs. placebo controls	various outcomes	.51
Chronic disease patient education	compliance and health	.52
Enrichment programs for gifted children	cognitive, creativity, affective outcomes	.55
Large (d = .80)		
Psychotherapy	various outcomes	.85
Meditation and relaxation techniques	blood pressure	.93
Positive reinforcement in the classroom	learning	1.17

outcomes such as functioning levels, blood pressure, and learning (*d* values range from .85 to 1.17).

It is important to keep in mind that "small," "medium," or "large" effect refers to the size of the effect, but not necessarily to its importance. For example, a new security screening procedure might lead to a small change in rates of detecting threats, but if this change translates into hundreds of lives saved at a small cost, then the effect might be judged to be both important and worth paying attention to.

As Fig. 1.2 suggests, when the true treatment effect is very small, it might be hard to accurately and consistently detect this effect in study samples. For example, aspirin can be useful in reducing heart attacks, but the effects are relatively small (d =.068; see, however, Rosenthal, 1993). As a result, studies of 20 or 30 patients taking aspirin or a placebo will not consistently detect the true and life-saving effects of this drug. Large sample studies, however, provide compelling evidence of the consistent effect of aspirin on heart attacks. On the other hand, if the effect is relatively large, then it is easy to detect, even with a relatively small sample. For example, cognitive ability has a strong influence on performance in school (*d* is about 1.10), and the effects of individual differences in cognitive ability are readily noticeable even in small samples of students.

Decision Criteria and Power

Finally, the standard or decision criteria used in hypothesis testing has a critical impact on statistical power. The standards used to test statistical hypotheses are usually set with a goal of minimizing Type I errors; alpha levels are usually set at .05, .01, or some other similarly low level, reflecting a strong bias against treating study outcomes that might be due to nothing more than sampling error as meaningful (Cowles & Davis, 1982). Setting a more lenient standard makes it easier to reject the null hypothesis, and although this can lead to Type I errors in those rare cases where the null is actually true, anything that makes it easier to reject the null hypothesis also increases the statistical power of the study.

As Fig. 1.1 shows, there is always a trade-off between Type I and Type II errors. Making it very difficult to reject the null hypothesis minimizes Type I errors (incorrect rejections), but also increases the number of Type II errors. That is, if the null is rarely rejected, sometimes sample results will be incorrectly dismissed as mere sampling error when they may in fact indicate the true effects of treatments. Numerous authors have noted that procedures to control or minimize Type I errors can substantially reduce statistical power, and may cause more problems (i.e., Type II errors) than they solve (J. Cohen, 1994; Sedlmeier & Gigerenzer, 1989).

Power Analysis and the General Linear Model

The following chapters describe a simple and general model for statistical power analysis. This model is based on the widely used *F* statistic. This statistic (and variations on the *F*) is used to test a wide range of statistical hypotheses in the context of the general linear model (J. Cohen & P. Cohen, 1983; Horton, 1978; Tatsuoka, 1993b). This statistical model provides the basis for correlation, multiple regression, analysis of variance, descriptive discriminant analysis, and all of the variations of these techniques. The general linear model subsumes a large proportion of the statistics that are widely used in the social sciences, and tying statistical power analysis to this model shows how the same simple set of techniques can be applied to an extraordinary range of statistical analyses.

STATISTICAL POWER OF RESEARCH IN THE SOCIAL AND BEHAVIORAL SCIENCES

Research in the social and behavioral sciences often shows shockingly low levels of power. Starting with J. Cohen's (1962) review of research published in the *Journal of Abnormal and Social Psychology*, studies in psychology, education, communication, journalism, and other related fields have routinely documented power in the range of from .20 to .50 for detecting small to medium treatment effects (Sedlmeier & Gigerenzer, 1989). Despite decades of warnings about the consequences of low levels of statistical power in the behavioral and social sciences, the level of power encountered in published studies is lower than .50 (Mone, Mueller, & Mauland, 1996). In other words, it is typical for studies in these areas to have less than a 50% chance of rejecting the null hypothesis. If researchers believe that the null hypothesis is virtually always wrong (i.e., that treatments have at least some effect, even if it is a very small one), then this means that at least one half of all studies in the social and behavioral sciences (perhaps as many as 80%) are likely to reach the wrong conclusion when testing the null hypothesis. This is even more startling and discouraging when they realize that these reviews have examined the statistical power of published research. Given the strong biases against publishing methodologically suspect studies or studies reporting null results, it is likely that the studies that survive the editorial review process are better than the norm, that they show stronger effects than similar unpublished studies, and that the statistical power of unpublished studies is even lower than the power of published studies.

Studies that do not reject the null hypothesis are often regarded by researchers as failures. The levels of power already reported suggest that "failure," defined in these terms, is quite common. If a treatment effect is small, and a study is designed with a power level of .20 (which is depressingly typical), it is four times as likely to fail (i.e., fail to reject the null) as to succeed. Power of .50 suggests that the outcome of the study is basically like the flip of a coin. The study is just as likely to fail as it is to succeed. It is likely that much of the apparent inconsistency in research findings is due to nothing more than inadequate power (Schmidt, 1992). If 100 studies are conducted, each with a power of .50, one half of them will and one half will not reject the null. Given the stark implications of low power, it is important to consider why research in the social and behavioral sciences is so often conducted in a way in which failure is more likely than success.

The most obvious possibility is that social scientists tend to study treatments, interventions, and so on. that have very small and unreliable effects. Until recently, this explanation was widely accepted, but the widespread use of meta-analysis in integrating scientific literature suggests that this is not the case. There is now ample evidence from literally hundred of analyses of thousands of individual studies that the treatments, interventions, and the like studied by behavioral and social scientists have substantial and meaningful effects (Haase, Waechter, & Solomon, 1982; J. E. Hunter & Hirsh,1987; Lipsey, 1990; Lipsey & Wilson, 1993; Schmitt, Gooding, Noe, & Kirsch, 1984); these effects are of a similar order of magnitude as many of the effects reported in the physical sciences (Hedges, 1987). A second possibility is that the decision criteria used to define statistical significance are too stringent. Several chapters herein argue that researchers are often too concerned with Type I errors and insufficiently concerned with statistical power. However, the use of overly stringent decision criteria is probably not the best explanation for low levels of statistical power.

The best explanation for the low levels of power observed in many areas of research is many studies use samples that are much too small to provide accurate and credible results. Researchers routinely use samples of 20, 50, or 75 observations to make inferences about population parameters. When sample results are unreliable, it is necessary to set some strict standard to distinguish real treatment effects from fluctuations in the data that are due to simple sampling error, and studies with these small samples often fail to reject null hypotheses, even when the population treatment effect is fairly large. On the other hand, very large samples will allow for rejection of the null hypothesis even when it is very nearly true (i.e., when the effect of treatments is very small). In fact, the effects of sample size on statistical power are so profound that it is tempting to conclude that a significance test is little more than a roundabout measure of how large the sample is. If the sample is sufficiently small, the null hypothesis would virtually never be rejected. If the sample is sufficiently large, the null hypothesis will virtually always be rejected.

USING POWER ANALYSIS

Statistical power analysis can be used for both planning and diagnosis. The most typical use of power analysis is in designing research studies. Power analysis can be used to determine how large a sample should be, or in deciding what criterion should be used to define statistical significance. Power analysis can also be used as a diagnostic tool, to determine whether a specific study has adequate power for specific purposes, or to identify the sort of effects that can be reliably detected in that study.

Because power is a function of the sensitivity of the study (which is essentially a function of *N*), the size of the effect in the population (*ES*), and the decision criterion used to determine statistical significance, it is possible to solve for any of the four values (i.e, power, *N*, *ES*, α), given the other three. However, none of these values is necessarily known in advance, although some values may be set by convention. The criterion for statistical significance (i.e., α) is often set at .05 or .01 by convention, but there is nothing sacred about these values. As is noted later, one important use of power analysis is in making decisions about what criteria should be used to describe a result as significant.

The effect size depends on the treatment, phenomenon, or variable being studied, and is usually not known in advance. Sample size is rarely set in advance, and *N* often depends on some combination of luck and resources on the part of the investigator. Actual power levels are rarely known, and it can be difficult to obtain sensible advice about how much power is necessary. It is important to understand how each of the parameters involved is determined when conducting a power analysis.

Determining the Effect Size

There is a built-in dilemma in power analysis. In order to determine the statistical power of a study, the effect size must be known. But if researchers already knew the exact strength of the effect of the particular treatment, intervention, and so forth, they would not need to do the study! The whole point of doing a study is to find out what effect the treatment has, and the true effect size in the population is unlikely to ever be known.

Statistical power analyses are always based on estimates of the effect size. In many areas of study, there is a substantial body of theory and empirical research that will provide a well-grounded estimate of the effect size. For example, there are literally hundreds of studies of the validity of cognitive ability tests as predictors of job performance (J. E. Hunter & Hirsch, 1987; Schmidt, 1992), and this literature suggests that the relation between test scores and performance is consistently strong (corrected correlations of about .50 are frequently seen). In order to estimate the statistical power of a study of the validity of a cognitive ability test, the results from this literature could be used to estimate the expected effect size. Even where there is not an extensive literature available, researchers can often use their experience with similar studies to realistically estimate effect sizes.

When there is no good basis for estimating effect sizes, power analyses can still be carried out by making a conservative estimate. A study that has adequate power to reliably detect small effects (e.g., a *d* of .20, a correlation of .10) will also have adequate power to detect larger effects. On the other hand, if researchers design their studies with the assumption that effects will be large, they might have insufficient power to detect small but important effects. Earlier, it was noted that the effects of taking aspirin on heart attacks are relatively small, but there is still a substantial payoff for taking the drug. If the initial research that led to the use of aspirin for this purpose had been conducted using small samples, then the researchers would have had little chance of detecting this life-saving effect.

Determining the Desired Level of Power

In determining desired levels of power, the researcher must weigh the risks of running studies without adequate power against the resources needed to attain high levels of power. You can always achieve high levels of power by using very large samples, but the time and expense required may not always justify the effort.

There are no hard and fast rules about how much power is enough, but there does seem to be consensus about two things. First, if at all possible, power should be above .50. When power drops below .50, the study is more likely to fail (i.e., it is unlikely to reject the null hypothesis) than to succeed. It is hard to justify designing studies in which failure is the most likely outcome. Second, power of .80 or above is usually judged to be adequate. The .80 convention is arbitrary (in the same way that significance criteria of .05 or .01 are arbitrary), but it seems to be widely accepted, and it can be rationally defended.

Power of .80 means that success (rejecting the null) is four times as likely as failure. It can be argued that some number other than four might represent a more acceptable level of risk (e.g., if power = .90, success is nine times as likely as failure), but it is often prohibitively difficult to achieve power much in excess of .80. For example, to have a power of .80 in detecting a small treatment effect (where the difference between treatment and control groups is $d = .20$), a total sample of about 775 subjects is needed. If researchers want power to be .95, they need about 1,300 subjects. Most power analyses specify .80 as the desired level of power to be achieved, and this convention seems to be widely accepted.

Applying Power Analysis

There are four ways to use power analysis: (a) in determining the sample size needed to achieve desired levels of power, (b) in determining the level of power in a study that is planned or has already been conducted, (c) in determining the size of effect that can be reliably detected by a particular study, and (d) in determining sensible criteria for statistical significance. The chapters that follow lay out the actual steps in doing a power analysis, but it is useful at this point to get a preview of the four potential applications of this method. Power analysis can be used in:

1. Determining sample size: Given a particular *ES*, significance criterion and a desired level of power, it is easy to solve for the sample size needed. For example, if researchers think the correlation between a new test and performance on the job is .30, and they want to have at least an 80% chance of rejecting the null hypothesis (with a significance criterion of .05), they need a sample of about 80 cases. When planning a study, routinely use power analysis to help make sensible decisions about the number of subjects needed.

2. Determining power levels: If *N*, *ES*, and the criterion for statistical significance are known, power analysis can be used to determine the level of power for that study. For example, if the difference between treatment and control groups is small (e.g., $d = .20$), there are 50 subjects in each group, and the significance criterion is $\alpha = .01$, then power will be only .05! Researchers should certainly expect that this study will fail to reject the null, and they might decide to change the design of their research considerably (e.g., use larger samples, more lenient criteria)

3. Determine *ES* levels: Researchers you can also determine what sort of effect could be reliably detected, given *N*, the desired level of power, and a. In the previous example, a study with 50 subjects in both the treatment and control groups would have power of .80 to detect a very large effect (approximately $d = .65$) with a .01 significance criterion, or a large effect ($d = .50$) with a .05 significance criterion.

4. Determine criteria for statistical significance: Given a specific effect, sample size, and power level, it is possible to determine the significance criterion. For example, if researchers expect a correlation

coefficient to be .30, N = 67, and they want power to equal or exceed .80, they will need to use a significance criterion of α = .10 rather than the more common .05 or .01.

HYPOTHESIS TESTS VERSUS CONFIDENCE INTERVALS

Null hypothesis testing has been criticized on a number of grounds (e.g., Schmidt, 1996), but perhaps the most persuasive critique is that they provide so little information. It is widely recognized that the use of confidence intervals and other methods of portraying levels of uncertainty about the outcomes of statistical procedures have many advantages over simple null hypothesis tests (Wilkinson & Task Force on Statistical Inference, 1999). For example, suppose a study is being done that examines the correlation between scores on an ability test and measures of performance in training. It find a correlation of r = .30, and on the basis of a null hypothesis test, it is decided that this value is significantly (e.g., at the .05 level) different from zero. That test tells researchers something, but it does not really tell them whether the finding that r = .30 represents a good or a poor estimate of the relation between ability and training performance. A confidence interval would provide that sort of information.

Staying with this example, suppose researchers estimate the amount of variability expected in correlations from studies like theirs, and conclude that a 95% confidence interval ranges from .05 to .55. This confidence interval would tell them exactly what they learned from the significance test (i.e., that they could be pretty sure the correlation between ability and training performance was not zero), but it would also tell then that r = .30 might not turn out to be a good estimate at all. Another researcher doing a similar study but using a larger sample might find a much smaller confidence interval, indicating a good deal more certainty about the generalizability of sample results.

As the previous paragraph above implies, most of the statements that can be made about statistical power also apply to confidence intervals. That is, if researchers design a study with low power, they will also find that it produces wide confidence intervals (i.e., that there is considerable uncertainty about the meaning of sample results). If they design studies to be sensitive and powerful, they will yield smaller confidence intervals. Thus, although the focus is on hypothesis tests, it is important to keep in mind that the same facets of the research design (N, the alpha level) that cause power to go up or down also cause confidence intervals to shrink or grow. A powerful study will not always yield precise results (e.g., power can be high in a poorly designed study that examines a treatment that has very strong effects), but in most instances, whatever is done to increase power will also lead to smaller confidence intervals and to more precision in sample statistics.

CONCLUSIONS

Power is defined as the probability that a study will reject the null hypothesis when it is in fact false. Studies with high statistical power are very likely to detect the effects of treatments, interventions, and so on, whereas studies with low power will often lead researchers to dismiss potentially important effects as sampling error. The statistical power of a test is a function of the size of the treatment effect in the population, the sample size, and the particular criteria used to define statistical significance. Although most discussions of power analysis are phrased in terms of traditional null hypothesis testing, where the hypothesis that treatments have no impact whatsoever is tested, this technique can be fruitfully applied to any method of statistical hypothesis testing.

Statistical power analysis has received less attention in the behavioral and social sciences than we think it deserves. It is still routine in many areas to run studies with disastrously low levels of power. Remember that statistical power analysis can be used to determine the number of subjects that should be included in a study, to estimate the likelihood that the study will reject the null hypothesis, to determine what sorts of effects can be reliably detected in a study, or to make rational decisions about the standards used to define statistical significance. Each of these applications of power analysis is taken up in the chapters that follow.

2

A Simple and General Model for Power Analysis

This chapter develops a simple approach to statistical power analysis that is based on the widely used *F* statistic. This statistic (or some transformation of *F*) is used to test statistical hypotheses in the general linear model (Horton, 1978; Tatsuoka, 1993b), a model that includes all of the variations of correlation and regression analysis (including multiple regression), analysis of variance and covariance (ANOVA and ANCOVA), *t* tests for differences in group means, tests of the hypothesis that the effect of treatments takes on a specific value, or a value different from zero. The great majority of the statistical tests used in the social and behavioral sciences can be treated as special cases of the general linear model.

This method is not the only approach to statistical power analysis. For example, in the most comprehensive work on power analysis, J. Cohen (1988) constructed power tables for a wide range of statistics and statistical applications, using separate effect size measures and power calculations for each class of statistics. Kramer and Thiemann (1987) derived a general model for statistical power analysis based on the intraclass correlation coefficient, and developed methods for expressing a wide range of test statistics in terms that were compatible with a single general table based on the intraclass *r*. Lipsey (1990) used the *t* test as a basis for estimating the statistical power of several statistical tests.

The idea of using the *F* distribution as the basis for a general system of statistical power analysis is hardly an original one; Pearson and Hartley (1951) proposed a similar model over 50 years ago. It is useful, however, to lay out the rationale for choosing the *F* distribution in some detail, because the family of statistics based on *F* have a number of characteristics that help to take some of the mystery out of power analysis.

Basing a model for statistical power analysis on the *F* statistic provides an optimal balance between applicability and familiarity. First, the *F* statistic is ubiquitous. This chapter and the next show how to transform a wide range of test statistics and effect size measures into *F* statistics, and how to use those *F* values in statistical power analysis. Because such a wide range of statistics can be transformed into *F* values, structuring power analysis around the *F* distribution allows coverage of a great deal of ground with a single set of tables.

Second, the approach developed here is flexible. Unlike other presentations of power analysis, this discussion does not limit itself to tests of the traditional null hypothesis (i.e., the hypothesis that treatments have no effect whatsoever). This particular type of test has been roundly criticized (J. Cohen, 1994; Meehl, 1978; Morrison & Henkel, 1970), and there is a need to move beyond such limited tests. Discussions of power analysis consider several methods of statistical hypothesis testing, and show how power analysis can be easily extended beyond the traditional framework in which the possibility that treatments have no effect whatsoever is tested. In particular, this discussion shows how the model developed here can be used to evaluate the power of minimum-effect hypothesis tests (i.e., tests of the hypothesis that the effects of treatments exceed some predetermined minimum level).

Recently, researchers have devoted considerable attention to alternatives to the traditional null hypothesis test (e.g., Murphy & Myors, 1999; Rouanet, 1996; Serlin & Lapsley, 1985, 1993), focusing in particular on tests of the hypothesis that the effect of treatments falls within or outside of some range of values. For example, Murphy and Myors (1999) discussed alternatives to tests of the traditional null hypothesis that involve specifying some range of effects that would be regarded as negligibly small, and then testing the hypothesis that the effect of treatments falls within this range (H_0) or falls above this range (H_1—i.e., the effects of treatments are so large that they can not reasonably be described as negligible). The *F* statistic is particularly well-suited to such tests. This statistic ranges in value from zero to infinity, with larger values accompanying stronger effects. As is shown in sections that follow, this property of the *F* statistic makes it easy to adapt familiar testing procedures to evaluate the hypothesis that effects exceed some minimum level, rather than simply evaluating the possibility that treatments have no effect at all.

Finally, the *F* distribution explicitly incorporates one of the key ideas of statistical power analysis (i.e., that the range of values that might be expected for a variety of test statistics depends in part on the size of the effect in the population). As is explained later, the notion of effect size is reflected very nicely in one of the three parameters that determines the distribution of the statistic *F* (i.e., the so-called noncentrality parameter).

THE GENERAL LINEAR MODEL, THE *F* STATISTIC, AND EFFECT SIZE

Before exploring the *F* distribution and its use in power analysis, it is useful to briefly describe the key ideas in applying the general linear model as a method of structuring statistical analyses, show how the *F* statistic is used in testing hypotheses according to this model, and describe a very general index of whether treatments, interventions, tests, and so on have large or small effects.

Suppose that 200 children are randomly assigned to one of two methods of reading instruction. Each child receives this instruction, either accompanied by audiovisual aids (e.g., computer software that "reads" to the child while showing pictures on a screen) or without the aids. At the end of the semester, each child's performance in reading is measured.

One way to structure research on the possible effects of reading instruction methods and/or audiovisual aids is to construct a mathematical model to explain why some children read well and others read poorly. This model might take a simple additive form:

$$y_{ijk} = \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk} \qquad [1]$$

where:

y_{ijk} = the score of child k, who received instruction method *i* and audio-visual aid *j*
α_i = the effect of the method of reading instruction
β_j = the effect of audiovisual aids
$\alpha\beta_{ij}$ = the effect of the interaction between method of instruction and audiovisual aids
e_{ijk} = the part of the child's score that cannot be explained by the treatments he or she received

In a linear model of this sort, researchers might reasonably ask several sorts of questions. First, it makes sense to ask whether the effect of a particular treatment or combination of treatments is large enough to allow them to rule out sampling error as an explanation for why people receiving one treatment obtain higher scores than people not receiving it. As is explained, the *F* statistic is well suited for this purpose.

Second, it makes sense to ask whether the effects of treatments, interventions, and so forth, are relatively large or relatively small. There are a variety of statistics that might be used in answering this question, but one very general approach is to estimate the percentage of variance in scores (*PV*) that is explained by the various effects included in the model. Regardless of the specific approach taken in statistical testing under the general linear model (e.g., analysis of variance or covariance, multiple regression, *t* tests), the goal of the model is always to explain variance in the dependent variable (i.e., to help researchers understand why some children obtained higher scores than others).

Linear models like the one shown divide the total variance in scores into that which can be explained by methods and treatment effects (i.e., the combined of effects of instruction audiovisual aids) and that which cannot be explained in terms of the treatments received by subjects. The percentage of variance (*PV*) associated with each effect in a linear model provides one very general measure of whether treatment effects are large or small (i.e., whether they account for a lot of the variance in the dependent variable or only a little), and the value of *PV* is closely linked to *F.*

There are a number of specific statistics used in estimating *PV,* notably η^2 (eta squared) and R^2, which are typically encountered in the contexts of the analysis of variance and multiple regression, respectively. The more general term *PV* is preferred, because it refers to a general index of the effects of treatments or interventions, not to any specific statistic or statistical approach. As we will shown later, estimates of *PV* are extremely useful in structuring statistical power analyses for virtually any of the specific applications of the general linear model.

THE *F* DISTRIBUTION AND POWER

Taking the ratio of two independent estimates of the variance in a population (e.g., $\sigma,^2_1$ and σ^2_2), this ratio is distributed as *F,* where:

$$F = \sigma^2_1 / \sigma^2_2 \text{, with degrees of freedom of } df_1 \text{ and } df_2 \qquad [2]$$

This F ratio can be used to test a wide range of statistical hypotheses (e.g., testing for the equality of means and variances). In the general linear model, the F statistic is used to test the null hypothesis (e.g., that the means are equal across treatments), by comparing some measure of the variability in scores due to the treatments to some measure of the variability in scores that might be expected as a result of simple sampling error. In its most general form, the F test in general linear models is:

$$F = \text{variability due to treatments/variability due to error} \qquad [3]$$

The distribution of the statistic F is complex, and depends in part on both the degrees of freedom of the hypothesis or effect being tested (df_{hyp}) and the degrees of freedom for the estimate of error used in the test (df_{err}). If treatments have no effect whatsoever, then the expected value of F is $df_{\text{err}}/(df_{\text{err}} - 2)$, which is very close to 1.0 for values of df_{err} much greater than 10. That is, if the traditional null hypothesis is true (i.e., treatments have no effect whatsoever), then expect to find F ratios of about 1.0. However, as noted earlier, also expect some variability, because of sampling error, in the F values actually obtained, even if the null hypothesis is literally true. Depending on the degrees of freedom (df_{hyp} and df_{err}), the F values that would be expected if the null hypothesis is true might cluster closely around 1.00, or they might vary considerably. The F tables shown in most statistics textbooks provide a sense of how much these values might vary strictly as a function of sampling error, given various combinations of df_{hyp} and df_{err}.

Finally, it is useful to note that the F and chi-squared distributions are closely related (the ratio of two chi-squared variables, each divided by its degrees of freedom, is distributed as F), and both distributions are special cases of a more general form (the gamma distribution).

The Noncentral *F*

Most familiar statistical tests are based on the central F distribution (i.e., the distribution of F statistics expected when the traditional null hypothesis is true). However, as noted earlier, interventions or treatments normally have at least some effect, and the distribution of F values that would be expected in any particular study is likely to take the form of a noncentral F distribution. The power of a statistical test is defined by the proportion of that noncentral F distribution that exceeds the critical value used to define statistical significance. The shape and range of values in the noncentral F distribution is a function of both the degrees of freedom (df_{hyp} and df_{err}) and the *noncentrality* parameter (λ). One way to think of the noncentrality parameter is that it is a function of just how wrong the traditional null hypothesis is. When $\lambda = 0$ (i.e., when the traditional null hypothesis is true), the noncentral F is identical to the central F that is tabled in most statistics texts.

The exact value of the noncentrality parameter is a function of both the effect size and the sensitivity of the statistical test (which is largely a function of the number of observations, N). For example, in a study where n subjects are randomly assigned to each of four treatment conditions, $\lambda = (n\Sigma(\mu_j - \mu)^2)/\sigma^2_e$, where μ_j and μ represent the population mean in treatment group j and the population mean over all four treatments, and σ^2_e represents the variance in scores due to sampling error. Horton (1978) noted that in many applications of the general linear model:

$$\lambda_{\text{est}} = SS_{\text{effect}} / MS_e \qquad [4]$$

where λ_{est} represents an estimate of the noncentrality parameter, SS_{effect} represents the sum of squares for the effect of interest, and MS_e represents the mean square error term used to test hypotheses about that effect. Using *PV* to designate the proportion of the total variance in the dependent variable explained by treatments (which means that 1 – *PV* refers to the proportion not explained), the noncentrality parameter can be estimated with the following equation:

$$\lambda_{\text{est}} = df_{\text{err}}\,[PV/(1 - PV)] \qquad [5]$$

Equations 4 and 5 provide a practical method for estimating the value noncentrality parameter in a wide range of applications of the general linear model.[1]

[1] Equations 4 and 5 are based on simple linear models, in which there is only one effect being tested, and the variance in scores is assumed to be due to either the effects of treatments or to error (e.g., this is the model that underlies the t test or the one-way analysis of variance). In more complex linear models, df_{err} does not necessarily refer to the degrees of freedom associated with variability in scores of individuals who receive the same treatment (within-cell variability in the one-way ANOVA model), and a more general form of Equation 5 ($\lambda_{\text{est}} = [(N - k) * (PV/(1 - PV))]$, where N represents the number of observations and k represents the total number of terms in the linear model) is needed. When N is large, Equation 5 yields very similar results to those of the more general form shown earlier.

The noncentrality parameter reflects the positive shift of the *F* distribution as the size of the effect in the population increases (Horton, 1978). For example, if *N* subjects are randomly assigned to one of *k* treatments, the mean of the noncentral *F* distribution is approximately $[(N - k/N - k - 2)/(1 + \lambda/(k - 1))]$, as compared to an approximate mean of 1.0 for the central *F* distribution. More concretely, assume that 100 subjects are assigned to one of four treatments. If the null hypothesis is true, then the expected value of *F* is approximately 1.0. However, if the effect of treatments is in fact large (e.g., *PV* = .25), expect to find *F* values substantially larger than 1.0 most of the time; here expect *F* values closer to 11.9 than to 1.0. In other words, if the true effect of treatments is large, expect to find large *F* values most of the time.

The larger the effect, the larger the noncentrality parameter, and the larger the expected value of *F*. The larger the *F*, the more likely that H_0 will be rejected. Therefore, all other things being equal, the more noncentrality (i.e., the larger the effect or the larger the *N*), the higher the power.

Using the Noncentral *F* Distribution to Assess Power

Chapter 1 laid out the three steps in conducting a statistical power analysis (i.e., determine critical value for significance, estimate effect size, estimate proportion of test statistics likely to exceed critical value). If we apply these steps are applied here, it follows that power analysis involves:

1. Deciding what value of *F* is needed to reject H_0. As discussed later in this chapter, this depends in part on the specific hypothesis being tested
2. Estimating the effect size and the degree of noncentrality. Estimates of *PV* allow the estimation of the noncentrality parameter of the *F* distribution.
3. Estimating the proportion of the noncentral *F* that lies above the critical *F* from Step 1.

In the following chapters, a simple method of conducting power analyses, based on the noncentral *F* distribution, is presented that may be used for a range of hypotheses. Appendix A discusses approaches to approximating the noncentral *F* distribution. Appendix B presents a table of *F* values obtained by estimating the noncentral *F* distribution over a range of df_{hyp}, df_{err}, and effect size values.[2] This table eliminates the difficulty of estimating noncentral *F* values, and more important, of directly computing power estimates for each statistical test performed. This table can be used to test both traditional and minimum-effect null hypotheses, and to estimate the statistical power of tests of both types of hypotheses.

A Note on Approximations

The methods described here are simple and general, but they are not always precise to the last decimal place. There are many statistical procedures that fall under the general linear model umbrella, and some specific applications of this model may present unique complications or distributional complexities. However, the methods developed here provide acceptably accurate approximations for the entire range of statistical tests covered under this general model. Approximations are particularly appropriate for statistical power analysis because virtually all applications of this technique are themselves approximations (because the exact value of the population effect size is rarely known for sure). That is, real-world applications of power analysis rarely depend on precise estimates, but rather depend on obtaining reasonable estimates to help inform decision-makers. Thus, a good approximation is quite acceptable.

For example, power analysis might be used to guide the selection of sample sizes or significance criteria in a planned study. Power analysis typically functions as a decision aid rather than as a precise forecasting technique, and it is rare that different decisions will be reached when exact versus approximate power values are known. That is, users of power analysis are likely to reach the same decisions if they know that power is about .80 in a particular study as they would reach if they knew that the power was precisely .815 for that study.

Statistical power analysis is an area where precision is not of sufficient value to justify the use of cumbersome methods in pursuit of the last decimal place. As a result, it is possible to use the general method developed here to approximate, with a high degree of precision, the more specific findings that are obtained when power analyses are tailored to specific analytic techniques (see J. Cohen,

[2] As noted in Appendix A, all tables presented in this book are based on the cumulative distribution function for the noncentral *F*. The precision of these tables has been assessed extensively, and the values presented in these tables allows reconstruction of with considerable accuracy, the power tables presented by J. Cohen (1988) and in other standard works.

1988, for discussions of power analysis for each of several types of statistical tests). This approach allows researchers to estimate statistical power for statistical tests in the general linear model, by translating specific test statistics or effect size measures into their equivalent *F* values.

TRANSLATING COMMON STATISTICS AND EFFECT SIZE MEASURES INTO *F*

The model developed here is expressed in terms of the *F* statistic, which is commonly reported in analysis of variance and multiple regression. However, many studies report their results in terms of something other than an *F* value. It is useful, therefore, to have at hand formulas for translating common statistics and effect size measures (e.g., *d*) into their *F* equivalents. Table 2.1 presents a set of formulas for doing just that.

TABLE 2.1

Translating Common Statistics Into *F* Equivalent Values

		Degrees of Freedom	
Statistic	*F equivalent*	df_{hyp}	df_{err}
t test for difference between means	$F(1, df_{err}) = t^2$	–	$N - 2$
Correlation coefficient	$F(1, df_{err}) = \frac{r^2 df_{err}}{1 - r^2}$	–	$N - 2$
Multiple R^2	$F(df_{hyp}, df_{err}) = \frac{R^2 / df_{hyp}}{(1 - R^2) / df_{err}}$	p	$N - p - 1$
Hierarchical regression	$F(df_{hyp}, df_{err}) = \frac{(R_F^2 - R_R^2) / df_{hyp}}{(1 - R_F^2) / df_{err}}$	k	$N - p - 1$
Chi-square (χ^2)	$F(df_{hyp}, df_{err}) = \chi2 / df_{hyp}$	df_{hyp}	∞
Standardized mean difference (d)	$F(1, df_{err}) = \frac{d^2 df_{err}}{4}$	–	$N - 2$
d (repeated measures)	$F(1, df_{err}) = \frac{d^2 df_{err}}{4\sqrt{1 - r_{ab}}}$	–	$N - 2$

For example, a study compared the effectiveness of two smoking cessation programs in a sample of 120 adults, who are randomly assigned to treatments, and used the t test to compare scores in these two groups. The reported t value of 2.48 would be equivalent to an F value of 6.15, with 1 and 118 degrees of freedom. A study (N = 112) that reported a multiple correlation of .45 between a set of four vocational interest tests and occupational choice measure would yield an F value of 9.06, with 3 and 107 degrees of freedom.

Hierarchical regression is used to determine the incremental contribution of k new predictor variables over and above the set of predictor variables already in an equation. For example, in a study with N = 250, two spatial ability tests were used to predict performance as an aircraft pilot; scores on these tests explained 14% of the variability in pilots' performance (i.e., R^2 = .14). Four tests measuring other cognitive abilities were added to the predictor battery, and this set of six tests explained 29% of the variance in performance (i.e., R^2 = .29). The F statistic that corresponds to this increase in R^2 is F(4, 243) = 12.83.

As Table 2.1 shows, χ^2 values can be translated in F equivalents. For example, if you found a χ^2 value of 24.56 with 6 degrees of freedom, the equivalent F value is 4.10, with df_{hyp} = 6 and df_{err} being infinite. Because the F table asymptotes as df_{err} grows larger, df_{err} = 10,000 (which is included on the F Table listed in Appendix B) represents an excellent approximation to infinite degrees of freedom for the error term.

Table 2.1 also includes the effect size measure d. This statistic is not commonly used in hypothesis testing per se, but it is widely used in describing the strength of effects, particularly when the scores of those treatment is compared to scores in a control group. Suppose that previous research suggests that the effect size d should be about .25. A study in which 102 subjects were randomly assigned to one of two treatments would be expected to yield an F value of 1.56, with 1 and 100 degrees of freedom. If d is .50, then an F value of 6.25 would be expected. If you used a repeated measures design was used (e.g., one in which scores of 101 subjects on a pretest and a posttest were compared, with a correlation of .60 between these scores), F would be 9.78.

In the aforementioned examples, the sample size has been included. The reason for this is that the value and the interpretation of the F statistic depends in part on the size of the sample (in particular, on the degrees of freedom for the error term, or df_{err}). In the preceding paragraph, a d value of .25 in a sample of 102 would yield F (1,100) = 1.56. In a sample of 227, the same d would translate into F (1,225) = 3.52. This reflects the fact that the same difference between means is easier to statistically detect when the sample (and therefore df_{err}) is large than when the sample is small. Small samples produce unstable and unreliable results, and in a small sample it can be hard to distinguish between true treatment effects and simple sampling error.

Finally, a note concerning terminology. The previous section and several sections that follow use the term "*F* equivalent." This term is meant to be explicit in recognizing that even when the results of a statistical test in the general linear model are reported in terms of some statistic other than *F* (e.g., *r*, *t*, *d*), it is nevertheless very often possible to the *F* value that is equivalent in meaning.

Transforming from *F* to *PV*

Table 2.1 shows how to transform commonly used statistics and effect size estimates into their equivalent *F* values. It may also be used to transform from *F* values to an equivalent effect size measure. For example, suppose that researchers used analysis of variance to analyze their data, reported a significant *F* value, but did not provide information about the strength of the effect. Equations 6 and 7 allow them to obtain an estimate of the proportion of variance in the dependent variable explained by the linear model (i.e., *PV*), given the value of *F*:

$$f^2 = (df_{\text{hyp}} * F) / df_{\text{err}} \qquad [6]$$

$$PV = f^2 / (1 + f^2) \qquad [7]$$

The f^2 in formulations 6 and 7 is the squared value of the effect size estimator f discussed in J. Cohen (1988). For example, if a study reports F (3,60) = 2.80, this implies that treatments account for about 12% of the variance in the dependent variable. As seen later in this chapter, formulas for transforming effect size estimates to their *F* equivalents, or *F* values to equivalent effect size estimates, are extremely useful in conducting power analyses.

Equations 6 and 7 can be combined into a single, simple formula for estimating *PV* on the basis of *F* and degrees of freedom:

$$PV = (df_{\text{hyp}} * F) / [(df_{\text{hyp}} * F) + df_{\text{error}}] \qquad [8]$$

As noted in chapter 5, this equation cannot not be used in complex, multifactor analysis of (ANOVA) variance designs, because the *F* statistic for any particular effect in a complex ANOVA model and its degrees of freedom do not contain all of the information needed to estimate *PV.* However, that chapter shows how to generalize a formula very similar to that shown in Equation 7 to complex ANOVA studies.

Nonparametric and Robust Statistics

The decision to anchor this discussion to statistical power analysis to the *F* distribution is driven primarily by the ubiquitousness of statistical tests based on that distribution. The *F* statistic can be used to test virtually any hypothesis that falls under the broad umbrella of the general linear model, but there are some important statistics that do not fall under this umbrella, and are not handled by the model developed and discussed here.

For example, a number of robust, or "trimmed," statistics have been developed in which outliers are removed from observed distributions prior to estimating standard errors and test statistics (Wilcox, 1992; Yuen, 1974). Trimming outliers can sometimes substantially reduce the effects of sampling error, and trimmed statistics can have more power than their normal theory equivalents (Wilcox, 1992). The power tables developed in this book are not fully appropriate for trimmed statistics, and can substantially underestimate the power of these statistics when applied in small samples.

A second family of statistics that is not easily accommodated using the model developed here are those statistics referred to as *nonparametric.* In general, nonparametric statistics do not make a priori assumptions about distributional forms, and tend to use little information about the observed distribution of data in constructing statistical tests. The conventional wisdom has long been that nonparametric tests have less power than their parametric equivalents (Siegel, 1956), but this is not always the case. Nonparametric tests can have more power than their parametric equivalents under a variety of circumstances, especially when conducting tests using distributions with heavy tails (Zimmerman & Zumbo, 1993). The methods developed here provide only gross approximations when used to assess the power of robust or nonparametric equivalents of standard statistical tests.

ALTERNATIVES TO THE TRADITIONAL NULL HYPOTHESIS

The traditional null hypothesis is that treatments, interventions, and so on have no effect. There are two advantages to testing this hypothesis: (a) It is easy (tests of this hypothesis are the standard fare for statistics courses, textbooks, data analysis packages, etc.), and (b) if the hypothesis that treatments have no effect is rejected, then researchers are left with the alternative that they have some effect. However, as noted earlier, this approach to statistical analysis has increasingly come under attack (J. Cohen, 1994; Meehl, 1978; Morrison & Henkel, 1970; Murphy, 1990; Schmidt, 1992, 1996. For discussions of the advantages of this approach, see Chow, 1988; Cortina & Dunlap, 1997; Hagen, 1997). The most general criticism of this approach is that nobody actually believes the null hypothesis. That is, it is rare to encounter treatments, interventions that have no effect whatsoever, which is the traditional null hypothesis. By definition, the traditional null hypothesis is usually quite likely to be false, which means that tests of this hypothesis are not necessarily meaningful (Murphy, 1990).

Second, the outcomes of tests of the traditional null hypothesis are routinely misinterpreted. As noted later, the outcomes of standard statistical tests probably say more about the power of the study than about the phenomenon being studied!

Why Is the Traditional Null Hypothesis Almost Always Wrong?

The traditional null hypothesis is formally wrong because it is a point hypothesis. That is, the hypothesis being tested is that the effect of treatments is exactly nil, even to the millionth decimal place or beyond. A treatment that has an obviously negligible effect will nevertheless lead to the rejection of the traditional null, as long as a test with sufficient statistical power is carried out.

The traditional null hypothesis represents a convenient abstraction, similar to the mythical "friction-less plane" encountered by freshmen in solving physics problems, in which small effects are treated as zero for the sake of simplicity. There are many real-world phenomena that seem to mirror the traditional null hypothesis, the most obvious being flipping a coin (see Fick, 1995, for examples that seem to show how the traditional null could be correct). However, even in studies that involved repeated flips of a fair coin, the traditional null is, by definition, wrong. The traditional null hypothesis is that there is no difference whatsoever in the probability of getting a head or a tail. In fact, it is impossible to mill a coin that is so precisely balanced that this will be true. Even if it is balanced to the billionth of the ounce, the traditional null is still wrong. Imbalance at the ten-billionth of the ounce will lead to rejection of traditional null if the coin is flipped a sufficient number of times.

The traditional null hypothesis is virtually always wrong because it is infinitely precise, whereas none of the real-world phenomena it is designed to test can possibly reach that level of precision. The same can be said for any point hypothesis (e.g., the hypothesis that the difference between two methods of reading instruction is exactly 6 points). The argument against the traditional null is not only a philosophical one; there are also abundant data to suggest that treatments in the social and behavioral sciences virtually always have at least some effect (Lipsey & Wilson, 1993; Murphy & Myors, 1999). In fact, it may not be possible to devise a real treatment that has no effect whatsoever; the hypothesis that treatments have no effect is so unlikely to be true that tests of this hypothesis are sometimes pointless (Murphy, 1990).

The fact that the traditional null hypothesis is virtually always wrong has important implications for thinking about Type I and Type II errors. If the null hypothesis is wrong, it is impossible to make a Type I error (i.e., to reject H_0 when it is true), and the only error that can occur is a Type II error (i.e., failing to reject H_0 when necessary). If researchers accept the position that the traditional null hypothesis is virtually never true, then it follows that they should be much more concerned with statistical power than with controlling of Type I errors.

Significance Tests are Routinely Misinterpreted

Another criticism of the traditional null hypothesis is that showing that a result is "significant" at the .05 level does not necessarily imply that it is important or it is especially likely to be replicated in a future study (J. Cohen, 1994). It merely shows that the particular findings being tested would probably not have been found if the true effect of treatments was zero. Unfortunately, researchers routinely misinterpret the results of these tests (J Cohen, 1994; Cowles, 1989; Greenwald, 1993). This is entirely understandable; most dictionary definitions of "significant" include synonyms such as "important" or "weighty." However, these tests do not directly assess the size or importance of treatment effects.

Tests of the traditional null hypothesis are more likely to tell about the sensitivity of the study than about the phenomenon being studied. With large samples, statistical tests of the traditional null hypothesis become so sensitive that they can detect the slightest difference between a sample result and the specific value that characterized the null hypothesis, even if this difference is negligibly small. With small samples, on the other hand, it is difficult to establish that anything has a statistically significant effect. The best way to get an appreciation of the limitations of traditional null hypothesis tests is to scan the tables in any power analysis book (J. Cohen, 1988). It becomes clear that, regardless of the true strength of the effect, the likelihood of rejecting the traditional null hypothesis is very small when samples are small, and is virtually certain when samples are large. Clearly, there is a need for approaches to significance testing that tell more about the phenomenon being studied than about the size of the samples. Alternatives to traditional null hypothesis tests are described next.

MINIMUM-EFFECT TESTS AS ALTERNATIVES TO TRADITIONAL NULL HYPOTHESIS TESTS

The arguments outlined have led some critics to call for abandoning null hypothesis testing altogether. But, rather than take this drastic step, it is better to reform the process. The problem with most null hypotheses is that the specific hypothesis being tested (i.e., that treatments have no effect whatsoever) is neither credible nor informative (Murphy, 1990); J. Cohen (1994) dismissed this approach as testing the "nil hypothesis." As noted earlier, there are several alternatives to testing the nil hypothesis, and all of these are a marked improvement over the standard procedure of testing the hypothesis that the effects of treatments is precisely zero. Serlin and Lapsley (1993) showed how to test the hypothesis that the effect of treatments falls within or outside of some range of values that is "good enough" to establish that one treatment is meaningfully better than the other. Rouanet (1996) demonstrated how Bayesian methods can be used to assert the importance or negligibility of treatment effects. Both of these methods allow for direct testing of credible and meaningful hypotheses.

The method proposed here is simply to shift the null distribution slightly. Instead of using a central distribution to represent the null hypothesis, use a noncentral one instead by setting the effect-size component of the noncentrality parameter to a small, nonzero effect. Then, when you rejecting H_0, researchers are rejecting the null hypothesis that their effect is negligibly small, rather than that it is precisely zero. As argued earlier, treatments are very unlikely to have a zero effect, but they may be negligible.

There is much to be learned by conducting tests of a substantive null hypothesis, such as the hypothesis that effects of treatments are negligibly small (e.g., they account for 1% or less of the variance in outcomes). In contrast to tests of the traditional null, tests of this sort are far from trivial (i.e., it is not known in advance whether H_0 is wrong), and they involve questions of real substantive interest. First, because they include some range of values rather than a single exact point under the "null umbrella," the results of these tests are not a foregone conclusion. Although it might be impossible to devise a treatment, intervention, and so on that had no effect whatsoever, there are any number of treatments whose effects fall somewhere between zero and whatever point is chosen to designate as a negligible effect. The possibility that your treatments will have a negligibly small effect is both real and meaningful (whereas the possibility that they will have no effect whatsoever is not), and researchers can learn something important about their treatments by testing this hypothesis.

As noted earlier, when testing the traditional null hypothesis, the outcome if can be guaranteed if the sample is sufficiently large. That is, with a sufficiently sensitive study, the statistical power of tests of the traditional null hypothesis can reach 1.00. The same is not true when testing the hypothesis that the effect of treatment exceeds some negligible minimum. Because there is a real possibility that the effect of treatments will not exceed that minimum, the upper bound of the power of these alternative tests will generally be lower than 1.00. No matter how good the study, researchers will never be certain in advance of the result of such a test. This may be regarded as a good thing. One major criticism of tests of the traditional null is that they are literally pointless (Murphy, 1990). Because it is known in advance that the traditional H_0 is wrong, nothing new is learned about H_0, regardless of the outcome of the significance test. Minimum-effect tests, on the other hand, can be informative, particularly if they have acceptable levels of statistical power. If there is a great deal of power and the test still fails to reject the hypothesis that a treatment effect is at best negligible, this would be regarded as strong evidence that the effect *is* negligible.

Testing the Hypothesis That Treatment Effects are Negligible

The best way to describe the process of testing a minimum-effect hypothesis is to compare it to the process used in testing the traditional null. The significance of the *F* statistic is usually assessed by comparing the value of the *F* obtained in a study to the value listed in an *F* table. The tabled values correspond to specific percentiles in the central *F* distribution. For example, if $df_{\text{hyp}} = 2$ and $df_{\text{err}} = 100$, the tabled values of *F* are 3.09 and 4.82, for $\alpha = .05$ and $\alpha = .01$, respectively. In other words, if the null hypothesis is true and there are 2 and 100 degrees of freedom, then researchers should expect to find *F* values of 3.09 or lower 95% of the time, and values of 4.82 or lower 99% of the time. If the *F* in the study is larger than these values, researchers will reject the null hypothesis and conclude that treatments probably do have some effect.

Tests of minimum-effect hypotheses proceed in exactly the same way, only using a different set of tabled *F* values (Murphy & Myors, 1999). The *F* tables found in the back of most statistics texts are based on the central *F* distribution, or the distribution of the *F* statistic that would be expected if the traditional null hypothesis were true. Tests of minimum-effect hypotheses are based on a noncentral F distribution. For example, suppose it is decided that treatments that account for 1% or less of the variance in outcomes have a "negligible" effect. It is then possible to estimate a noncentrality parameter (based on $PV = .01$), and to estimate the corresponding noncentral F distribution for testing the hypothesis that treatment effects are at best negligible. If $PV = .01$, $df_{\text{hyp}} = 2$, and $df_{\text{err}} = 100$, then 95% of the values in this noncentral F distribution will fall at or below 4.49; 99% of the values in this distribution will fall at or below 6.76 (as noted later, F values for testing minimum-effect hypotheses are listed in Appendix B). In other words, if the observed F in the study was greater than 4.49, researchers could be confident ($\alpha = .05$) in rejecting the hypothesis that treatments accounted for 1% or less of the variance. Later in this chapter, we discuss standards that might be used in designating effects as negligible are discussed.

Minimum-effect hypotheses involve specifying a whole range of values as negligible. In the previous example, effects that account for 1% or less of the variance in the population were designated as negligible effects, and if researchers can reject the hypothesis that the effects are negligibly small, they are left with the alternative hypothesis that they are not negligibly small (i.e., that treatment effects are large enough to care about). But how does a single critical F value allows for testing of a whole range of null possibilities? Remember that one characteristic of the F statistic is that it ranges from zero to infinity, with larger F values indicating larger effects. Therefore, if researchers can be 95% confident that the observed F is larger than the F that would have been obtained if treatments accounted for 1% of the variance, then they can also be at least 95% confident that the observed F would be larger than that which would have been obtained for any PV value between .00 and .01. If the observed F is larger than the F values expected 95% of the time when $PV = .01$, then it must also be larger than 95% of the values expected for any smaller PV value.

An Example

Suppose 125 subjects are randomly assigned to one of five treatments. It is found that $F(4, 120) = 2.50$, and in this sample, treatments account for 7.6% of the variance in the dependent variable. The F is large enough to allow for the rejection of the traditional null hypothesis ($\alpha = .05$), but there is concern that the true effect of treatments might be negligibly small. Suppose also that in this context, treatments that account for less than 1% of the variance in the population have effects considered to be negligible. To test the hypothesis that the effects observed here came from a population in which the true effect of treatments is negligibly small, consult the noncentral F distribution with $df_{\text{hyp}} = 4$, $df_{\text{err}} = 120$, and $\lambda = 1.21$ (i.e., $\lambda_{\text{est}} = [120 * .01]/[1 - .01]$). It would be found that 95% of the values in this distribution are 3.13 or lower. The obtained F is 2.50, which is smaller than this critical value, and means that this null hypothesis cannot be rejected. That is, although researchers can reject the hypothesis that treatments have no effect whatsoever (i.e., the traditional null), they cannot reject the hypothesis that the effects of treatments are negligibly small (i.e., a minimum-effect hypothesis that treatments account for less than 1% of the variance in outcomes).

Defining a Minimum Effect

The main advantage of the traditional null hypothesis is that it is simple and objective. If researchers reject the hypothesis that treatments have no effect, they are left with the alternative that they have some effect. On the other hand, testing minimum-effect hypotheses requires value judgments, and requires that some consensus be reached in a particular field of inquiry. For example, the definition of a negligible effect might reasonably vary across areas, and there may be no set convention for defining which effects are so small that they can be effectively ignored and those that cannot be ignored. An effect that looks trivially small in one discipline might look pretty good in another. However, it is possible to offer some broad principles for determining when effects are likely to be judged negligible.

First, the importance of an effect should depend substantially on the particular dependent variables involved. For example, in medical research it is common for relatively small effects (in terms of the percentage of variance explained) to be viewed as meaningful and important (Rosenthal, 1993). One reason is that the dependent variables in these studies often include quality of life, and even survival. A small percentage of variance might translate into many lives saved.

Second, decisions about what effects should be labeled as negligible might depend on the relative likelihood and relative seriousness of Type I versus Type II errors in a particular area. As noted in a later section, the power of statistical tests in the general linear model decreases as the definition of a negligible effect expands. In any particular study, power is higher for testing the traditional null hypothesis that treatments have no effect than for testing the hypothesis that they account for 1% or less of the variance in outcomes, and higher for tests of the hypothesis that treatments account for 1% or less of the variance than for the hypothesis that treatments account for 5% or less of the variance in outcomes. If Type II errors are seen as particularly serious in a particular area of research, then it might make sense to choose a very low figure as the definition of a negligible effect.

On the other hand, there are many areas of inquiry in which numerous well-validated treatments are already available (See Lipsey & Wilson, 1993, for a review of numerous meta-analyses of treatment effects), and in these areas, it might make sense to "set a higher bar" by testing a more demanding hypothesis. For example, in the area of cognitive ability testing (where the criterion is some measure of performance on the job or in the classroom), it is common to find that tests account for from 20% to 25% of the variance in the criterion (J. E. Hunter & R. F. Hunter, 1984; J. E. Hunter & H. R. Hirsch, 1987). Tests of the traditional null hypothesis (i.e., that tests have no relation whatsoever to these criteria) are relatively easy to reject; if $\rho^2 = .25$, a study with $N = 28$ will have power of .80 for rejecting the traditional null hypothesis (J. Cohen, 1988). Similarly, the hypothesis that tests account for 1% or less of the variance in these criteria is easy to reject; if $\rho^2 = .25$, a study with $N = 31$ will have power of .80 for rejecting this minimum-effect hypothesis (Murphy & Myors, 1999). In this context, it might make sense to define a negligible relation as one in which tests accounted for 10% or less of the variance in these criteria.

Utility analysis has been used to help determine whether particular treatments have effects that are large enough to warrant attention (Landy, Farr, & Jacobs, 1982; Schmidt, J. E. Hunter, McKenzie, & Muldrow, 1979; Schmidt, Mack, & J. E. Hunter, 1984). Utility equations suggest another important parameter that is likely to affect the decision of what represents a negligible versus a meaningful effect (i.e., the standard deviation of the dependent variable, or SD_y). When there is substantial and meaningful variance in the outcome variable of interest, a treatment that accounts for a relatively small percentage of variance might nevertheless lead to practical benefits that far exceed the costs of the treatment.

For example, suppose a training program costs $1,000 per person to administer, and it is proposed as a method of improving performance in a setting where the current SD_y (i.e., the standard deviation of performance) is $10,000. If the effects of training account for less that 1% of the variation in job performance, then it might be concluded that the projected cost of training will exceed the projected benefit [based on the equation $\Delta U = r_{xy} \cdot SD_y - C$, where ΔU is the projected overall benefit, r_{xy} is the relation between the training and the criterion (in this case, $PV = .01$ translates into $r_{xy} = .10$), and C represents the cost; see Landy, Farr & Jacobs, 1982], which suggests that $PV = .01$ represents a sensible definition of the minimum effect needed to label training effects as nontrivial.

Many of the examples presented in this chapter and chapters that follow use conventions similar to those described in J. Cohen (1988), describing treatments that have less than 1% of the variance as having small effects, and those that account for less than 5% of the variance in outcomes as having small to medium effects (these conventions are discussed in detail later in this chapter). Many tables presented in this book are arranged around these particular conventions. However, it is critical to note that the decision of what represents a negligible effect is one that is likely to vary across research areas, and there will be many cases in which these particular conventions do not apply. Appendix A presents the information needed to determine critical *F* values for minimum-effect tests that employ some other operational definition of negligible, and researchers are irged to carefully consider their reasons for choosing any particular value as a definition of the minimum effect of interest.

Power of Minimum-Effect Tests

As this example suggests, expect less power when testing the hypothesis that the effect in the study exceeds some minimum value than when testing the hypothesis that it is exactly zero. Switching from a central to a noncentral distribution as the basis of a null hypothesis necessarily increases the criterion for significance, thereby reducing the proportion of the H_1 distribution that lies above this value (i.e., power). The traditional hypothesis that treatments have no effect whatsoever is by definition usually wrong (Murphy, 1990), and is therefore relatively easy to reject. If the sample is large enough, researchers will always reject the hypothesis that treatments have no effect, even if the true effect of treatments is extremely negligible. Tests of the hypothesis that treatment effects exceed whatever lower bound that is used to define negligible are more demanding than tests of the traditional null, in part because there is always a chance that the effects are negligible. Therefore, there is no guarantee that the hypothesis will be rejected, no matter how sensitive the study. However, as noted in several of the chapters that follow, the lower power of substantive tests is easily offset by the fact that these tests tell something meaningful, regardless of how large the sample size or how sensitive it is possible to make the study.

ANALYTIC AND TABULAR METHODS OF POWER ANALYSIS

There are several methods available to carry out statistical power analyses. First, it is possible to compute either the approximate or exact level of power in any particular case. Analytic methods of power analysis are discussed next. Second, it is possible to develop tables or graphs that provide good approximations of the statistical power of studies under a wide range of conditions. Our preference is to work with tables, in part because the type of graphs needed to plot a variable (i.e., power) as a function of three other variables (i.e., *N*, a, and *ES*) are complicated and difficult to use. As is discussed later, it is possible to generate tables that neatly integrate information about traditional and minimum-effect hypothesis tests with information about statistical power.

Analytic Methods

The previous section demonstrated the analytic approach to power analysis using either an approximation to or a relatively exact calculation of the noncentral *F* distribution. It is possible to apply this approach to calculate the power of any test that can be framed in terms of the familiar *F* statistic. All this requires is a calculator and a fairly detailed table of the widely used central *F* distribution (a table that only gives critical values at the .05 and .01 levels will not always have the information needed to calculate power). Appendix A discusses several methods of estimating values in the noncentral *F* distribution.

Although this analytic method is both precise and flexible, it is also relatively cumbersome and time consuming. That is, the direct computation of statistical power involves determining some standard for statistical or practical significance (e.g., setting a minimum value for a negligible effect), estimating the noncentral *F* distribution that applies, and determining the proportion of that distribution that lies above the standard. Even with a relatively powerful computer, the processes can be time consuming, and may be daunting to many consumers of power analysis. A more user-friendly approach is to develop tables that contain the essential information needed to estimate statistical power.

Power Tables

A number of excellent books present extensive tables describing the statistical power of numerous tests; J. Cohen (1988) is the most complete source currently available. The approach described here is simpler (although it provides a bit less information), but considerably more compact. All of the information needed to do most significance tests and power analyses for statistical tests in the general linear model is presented in Appendix B, which contains what is called the "One-Stop *F* Table." This is called a "one-stop" table because each cell contains the information needed for conducting traditional significance tests, conducting power analyses at various key levels of power, testing the hypothesis that the effect in a study exceeds various criteria used to define negligibly small or small to moderate effects, and estimating power for these minimum-effect tests.

USING THE ONE-STOP *F* TABLE

Each cell in the One-Stop *F* Table contains 12 pieces of information. The first four values in each cell are used for testing significance and estimating power for traditional null hypothesis tests. The next eight values in each cell are used for testing significance and estimating power when testing the hypothesis that treatment effects are negligible, using two different operational definitions of a negligible effect (i.e., treatments account for 1% or less of the variance, or they account for 5% or less of the variance).

To illustrate the use of this table, consider another study. Fifty-four subjects are randomly assigned to one of four treatments. The treatment being studied is one believed to have at least a moderate effect (e.g., researchers expect treatments to account for about 15% of the variance). It is found that F (3, 50) = 3.50, and treatments account for 17.3% of the variance in the sample.

Traditional Null Hypothesis Tests

The first four values in each cell of the One-Stop *F* table are used for hypothesis testing and power analysis when testing the traditional null hypothesis (i.e., that treatments have no effect). The first value in each cell of this table represents the critical *F* for the traditional null hypothesis significance test, with an alpha level of .05. In the previous example, $df_{\text{hyp}} = 3$ and $df_{\text{err}} = 50$, which yields a critical *F* of 2.79. If the observed *F* value in the study exceeds 2.79 (as was the case here), H_0 would be rejected. The second value in each cell is the critical *F* for the traditional null hypothesis test at the .01 level of significance (an *F* greater than or equal to 4.20 would be needed to reject the traditional null hypothesis at the .01 level). The *F* of 3.50 found in the study would allow rejection of the traditional null with $\alpha = .05$, but not with $\alpha = .01$.

The next two values in each cell are *F* equivalents of the effect size values needed to obtain particular levels of power (given an a level of 0.05 and the specified df_{hyp} and df_{err}). The values in the table are 1.99 and 3.88, respectively, for power levels of .50 and .80, respectively. If Equations 5 and 6 are used to transform these values into equivalent

PV values, values of .11 and .19 are found, respectively. That is, there would be a power of .50 for rejecting the traditional null if treatments accounted for about 11% of the variance in the population, and power of .80 if treatments accounted for about 19% of the variance. In fact, it would be expected that the population effect is somewhere between these two figures (i.e., treatments account for about 15% of the variance), which implies that the power of the study is in the range from .50 to .80. As shown later in this chapter, it is easy to estimate where in this range the power of the study actually falls (in this example, power is approximately .65).

Testing Minimum-Effect·Hypotheses (*PV* = .01)

Rather than testing the hypothesis that treatments have no effect whatsoever, researchers might want to test the hypothesis that treatment effects are so small that they account for less than 1% of the variance in outcomes. If this *PV* value represents a sensible definition of a negligible effect in a particular area of research, and they test and reject this hypothesis they can be confident that effects are not negligibly small.

The fifth and sixth values in each cell of the One-Stop *F* Table are the critical *F* values that would be needed in order to achieve significance (at the .05 and .01 levels, respectively) when testing the hypothesis that treatments account for 1% or less of the variance in outcomes. With $df_{\text{hyp}} = 3$ and $df_{\text{err}} = 50$, an *F* of 3.24 would be needed to reject (with a = .05) the hypothesis that the effect being studied was negligibly small (defined as accounting for 1% or less of the variance in outcomes). The obtained *F* was 3.50, meaning that both the traditional null hypothesis and the hypothesis that the effect is negligibly small can be rejected. However, if researchers use a .01 alpha level, their obtained *F* will be smaller than the tabled *F* value of 4.85. With an alpha level of .01, they will not be able to reject either the traditional null or the hypothesis that the observed effect is negligibly small.

The seventh and eighth values in each cell are *F* equivalents of the effect size values needed to obtain particular levels of power (given an a level of .05 and the specified df_{hyp} and df_{err}) for testing this substantive hypothesis. The values in the table are 2.46 and 4.48 for power levels of .50 and .80, respectively. This translates into *PV* values of .13 and .21, respectively. That is, if treatments accounted for 13% of the variance in the population, a study with $df_{\text{hyp}} = 3$, $df_{\text{err}} = 50$, and $\alpha = .05$ would have power of about .50. If treatments account for 21% of the variance in the population, then power would be approximately .80.

In this study, it was expected that treatments would account for about 15% of the variance. This analysis suggests that the power for testing the hypothesis that the population effect is negligibly small is somewhere between .50 and .80. Chapter 3 shows how to determine where in this range the power of this study lies.

Testing Minimum-Effect Hypotheses (*PV* = .05)

There are many treatments that routinely demonstrate moderate to large effects. For example, well-developed cognitive ability tests allow for the prediction of performance in school and in many jobs with a relatively high degree of success (correlations in the .30 – .50 range are common). Rather than testing the hypothesis that tests have no relation whatsoever with these criteria (i.e., the traditional null), or even that treatments account for 1% or less of the variance in outcomes, it might make sense to test a more challenging hypothesis (i.e., that the effect of this particular treatment is at least small to moderate in size). For reasons that are explained in the section that follows, there are many contexts in which it is useful to test the hypothesis that treatments explain 5% or less of the variance in the population. If researchers can test and reject the hypothesis that effects explain 5% or less of the variance in outcomes, they are left with the alternative that treatments explain more than 5% of the variance. In most contexts, effects this large are likely to be treated as meaningful, even if smaller effects (e.g., those accounting for 1% of the variance) are not.

The 9th and 10th values in each cell of the One-Stop *F* Table are the critical *F* values needed in order to achieve significance (at the .05 and .01 levels, respectively) when testing the hypothesis that treatments account for 5% or less of the variance in outcomes. With $df_{\text{hyp}} = 3$ and $df_{\text{err}} = 50$, an *F* of 4.84 is necessary to reject (with $\alpha = .05$) the hypothesis that the effect being studied was no greater than small to moderate in size. The obtained *F* was 3.50, meaning that it is impossible to reject this hypothesis (if a = .01, you would need *F* = 6.98 to reject this null). In other words, the findings of the study are not sufficiently strong to reject the possibility that the real population effect is smaller than what has been defined as the small to moderate range (i.e., *PV* = .05 or lower).

The 11th and 12th values in each cell are *F* equivalents of the effect size values needed to obtain particular levels of power (given an a level of 0.05 and the specified df_{hyp} and df_{err}) for tests of this minimum-effect hypothesis. The values in the table are 4.08 and 6.55 for power levels of .50 and .80, respectively. This translates into *PV* values of .20 of and .28, respectively. The treatments were expected to account for about 15% of the variance. This analysis suggests that the power for testing the hypothesis that this population effect is small to moderate (i.e., that it accounts for more than 5% of the variance in outcomes) is well below .50. In other words, this particular study does not have sufficient power for a credible test of this hypothesis. With this combination of *ES*, sample size, and significance criteria, a significance test is almost like the flip of a coin.

To review what this table reveals about the study, in which it was found that $F(3, 50) = 3.50$, and that treatments account for 17% of the variance in the sample, both the traditional null and the hypothesis that treatment effects are negligibly small can be rejected with an alpha of .05; with alpha of .01, neither hypothesis can be rejected. With an alpha level of .05, the table suggests that the power of tests of the traditional null hypothesis is between .50 and .80 (the actual power level is about .65). This study also has power slightly greater than .50 for testing the hypothesis that treatment effects are negligible (i.e., that the population *PV* is .01 or less). However, the hypothesis that treatments explain 5% or less of the variance in the population cannot be rejected.

Note that the One-Stop *F* Table does not contain a set of rows corresponding to relatively large effects (i.e., effects accounting for more than 5% of the variance in outcomes). As noted in chapter 1, when the effect of treatments is known or thought to be large, there is often no point to conducting the research. Large effects are usually so obvious that a study confirming their existence is unlikely to make much of a contribution. More to the point, when the effects under consideration are large, statistical power is unlikely to be a problem unless samples are extremely small. When samples are this small, there are problems that are much more severe than statistical power (e.g., lack of generalizability), and power analyses for large effects strike us as very limited in value.

Interpolating Between Tabled Values

Like all *F* tables, the One-Stop *F* Table is incomplete in that it does not table all possible values of df_{hyp} and df_{err}. Fortunately, relatively good approximations to all of the values in this table can be obtained by linear interpolation. For example, the table includes df_{err} values of 50 and 60. If researchers wanted to find appropriate *F* values for $df_{\text{err}} = 55$, these would lie about halfway between the values for $df_{\text{err}} = 50$ and $df_{\text{err}} = 60$. Thus, the approximate *F* needed to reject the traditional null hypothesis ($\alpha = .05$) with $df_{\text{hyp}} = 2$, $df_{\text{err}} = 55$ would be 3.165 (i.e., halfway between 3.15 and 3.18). Similarly, if $df_{\text{err}} = 48$, the appropriate *F* values could be estimated by computing the value that *F* that was 80% of the distance between the tabled *F* for $df_{\text{err}} = 40$ and the tabled *F* for $df_{\text{err}} = 50$. In general, the value of the interpolated *F* can be obtained using the following equation:

$$F_{\text{Interpolated}} = F_{\text{Below}} + \left(\left(\frac{\text{dfe}_{\text{int}} - \text{dfe}_{\text{below}}}{\text{dfe}_{\text{above}} - \text{dfe}_{\text{below}}} \right) \left(F_{\text{Above}} - F_{\text{Below}} \right) \right) \qquad [9]$$

where:

F_{below} = Tabled F below the value to be calculated
F_{above} = Tabled F above the value to be calculated
$df_{\text{below}} = df_{\text{err}}$ for tabled F below the value to be calculated
$df_{\text{above}} = df_{\text{err}}$ for tabled F above the value to be calculated
$df_{\text{int}} = df_{\text{err}}$ for F value to be calculated

It is important to keep in mind that linear interpolation will yield approximate values only. For the purposes of statistical power analyses, these interpolations will virtually always be sufficiently accurate to help in making sensible decisions about the design of studies, the choice of criteria for defining statistical significance, and so forth.

A second application of linear interpolation is likely to be even more useful. The table includes F equivalents for the effect size values needed to obtain power levels of .50 and .80, respectively. In the example, where $df_{\text{hyp}} = 3$ and $df_{\text{err}} = 50$, F values of 1.99 and 3.88 are equivalent to the population *PV* values needed to obtain power levels of .50 or .80 in a study like this [i.e., you would achieve these levels of power if treatments truly accounted for 11% of the variance (yielding power of .50) or 19% of the variance (yielding power of .80)]. Here, treatments were expected to account for 15% of the variance. If this figure is translated into its F equivalent (using formulas presented in Table 2.1), an F of 2.94 is obtained. A linear interpolation is used to estimate the power of the study, using a formula that closely parallels the formula used to interpolate F values:

$$\text{Power}_{\text{Interpolated}} = .50 + \left(\left(\frac{F_{\text{hypothesized}} - F_{50}}{F_{.80} - F_{.50}} \right) * .30 \right) \qquad [10]$$

where:

$F_{\text{hypothesized}}$ = F equivalent for hypothesized size of the effect
$F_{.50}$ = F equivalent of the *PV* needed to obtain power of .50 (α = .05)
$F_{.80}$ = F equivalent of the *PV* needed to obtain power of .80 (α = .05)

In the example, 54 subjects are assigned to one of four treatments (df_{hyp} = 3, df_{err} = 50), where treatments are expected to account for 15% of the variance, and where a = .05. In this example, $F_{hypothesized}$ = 2.94, $F_{.50}$ = 1.99, and $F_{.80}$ = 3.84. Using Equation 8, the estimate of power of this study for rejecting the traditional null hypothesis is .65. If the hypothesis tested was that treatments accounted for 1% or less of the variance, then $F_{.50}$ = 2.46, and $F_{.80}$ = 4.48, and power is estimated to be .57.

It is also easy to develop linear interpolation formulas that work directly from *PV* values, rather than working with their *F* equivalents. For example, suppose that researchers are carrying out a study in which 104 subjects are randomly assigned to one of four treatments, yielding df_{hyp} = 3, df_{err} = 100. The One-Stop *PV* Table shown in Appendix C tells them that if treatments account for 5.5% of the variance in the population, they will have power of .50 for testing the nil hypothesis (i.e., $PV_{.5}$ = .055). If treatments account for 10.1% of the variance in the population, they will have power of .80 for testing the nil hypothesis (i.e., $PV_{.8}$ = .101). They hypothesize that treatments in fact account for 7.5% of the variance (i.e., $PV_{hypothesized}$ = .075). Equation 11 can be used to estimate the power of the study:

$$\text{Power}_{\text{Interpolated}} = .50 + \left(\left(\frac{\text{PV}_{\text{hypothesized}} - \text{PV}_{.50}}{\text{PV}_{.80} - \text{PV}_{.50}} \right) * .30 \right) \qquad [11]$$

According to Equation 11, this study should be expected to have a power of .63 for testing the traditional null hypothesis.

Extrapolating outside of the tabled range of .50 – .80 is not recommended, and indeed estimating actual power levels when power is known to fall outside of these bounds does not seem to make much difference. If power is above .80, then it is unlikely that there will be different conclusions than if it was exactly .80, regardless of the outcome of a study. If power is below .50, there are more serious problems than the precise estimation of power.

THE ONE-STOP *PV* TABLE

Appendix C presents a table that parallels the One-Stop *F* Table, this time expressing the tabled values in terms of *PV* statistics rather than *F*. So, for example, suppose a study is conducted in which 93 subjects are randomly assigned to one of three treatments. Data from this study would normally be analyzed using ANOVA, with df_{hyp} = 2 and df_{err} = 90. The One-Stop *F* Table shown in Appendix B would give you the *F* values needed for doing significance tests (either traditional on minimum-effect tests) and for estimating the *F* equivalents needed to attain various levels of power. The One-Stop *PV* Table shown in Appendix C displays the same information expressed in *PV* values. For example, this table shows that to reject the traditional null hypothesis (with $\alpha = .05$), treatments will need to account for at least 6.4% of the variance in outcomes. If you set a more stringent alpha (e.g., $\alpha = .01$) is set, treatments will need to account for 9.7% of the variance to reject the traditional null.

If a minimum-effect hypotheses is being tested, larger *PV* values will be needed to reject the hypothesis that the effect of treatments is negligibly small. For example, if the null hypothesis is that treatments account for less than 1% of the variance in outcomes and $\alpha = .05$, a $PV = .089$ is necessary to reject this null. With $\alpha = .01$, a $PV = .129$ is needed to reject this minimum-effect null.

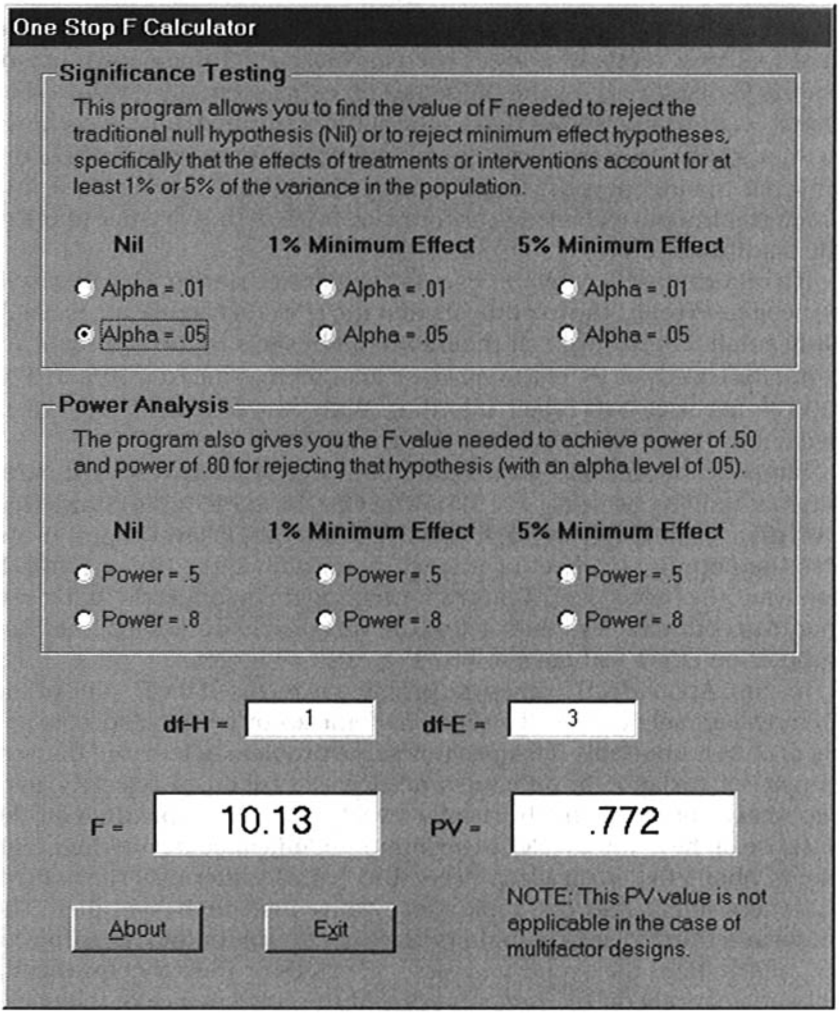
Suppose researchers want to obtain power of .80 in testing the traditional null hypothesis ($\alpha = .05$). The One-Stop *PV* Table shows that with $df_{\text{hyp}} = 2$ and $df_{\text{err}} = 90$, they will not attain this level of power unless the population effect is at least moderately large ($PV = .099$). If you want the same level of power for testing the hypothesis that treatments account for less than 1% of the variance in outcomes, then the population effect will have to be $PV = .128$ or larger.

In sum, Appendix B frames the problem in terms of the *F* value or the *F* equivalent needed to test various hypotheses or to reach specific levels of power. Appendix C frames this same problem in terms of the percentage of variance in outcomes needed to explain to reject various hypotheses or to establish various levels of power. Depending on the purpose of the analysis (e.g., determining significance vs. making statements about the effect sizes needed to satisfy different criteria), researchers might find one table more convenient than the other. The important thing to keep in mind is that both tables contain the same basic information; the choice to look for *F* values or their *PV* equivalents depends entirely on the preferences and the convenience of the user.

THE ONE-STOP *F* CALCULATOR

Finally, it is often easier and more convenient to use a computer program rather than a set of tables to estimate statistical power and to evaluate the effects of changes in the design of a study on the levels of power likely to be found. The One-Stop *F* Calculator program distributed with this book serves that purpose.

Figure 2.1 illustrates the screen that opens when the One-Stop *F* Calculator is installed on the computer (simple installation instructions are included on the CD that contains the program files). The program asks users to enter two quantities, the degrees of freedom for the



One Stop F Calculator

Significance Testing

This program allows you to find the value of F needed to reject the traditional null hypothesis (Nil) or to reject minimum effect hypotheses, specifically that the effects of treatments or interventions account for at least 1% or 5% of the variance in the population.

Nil	1% Minimum Effect	5% Minimum Effect
○ Alpha = .01	○ Alpha = .01	○ Alpha = .01
● Alpha = .05	○ Alpha = .05	○ Alpha = .05

Power Analysis

The program also gives you the F value needed to achieve power of .50 and power of .80 for rejecting that hypothesis (with an alpha level of .05).

Nil	1% Minimum Effect	5% Minimum Effect
○ Power = .5	○ Power = .5	○ Power = .5
○ Power = .8	○ Power = .8	○ Power = .8

df-H = 1 df-E = 3

F = 10.13 PV = .772

NOTE: This PV value is not applicable in the case of multifactor designs.

FIG. 2.1. The One-Stop *F* calculator.

hypothesis being tested (enter this in the box labeled *df*-H) and the the degrees of freedom for error (enter this in the box labeled *df*-E). Then users decide whether they want critical values for significance testing (either traditional nil hypothesis tests or minimum-effects tests, for $\alpha = .05$ or .01) or for estimating power (again for either nil or minimum-effects tests, for power levels of .50 or .80). The program gives both the *F* value and the equivalent *PV* value for the desired test.

Figure 2.1 illustrates a study that attempts to use the *t* test to compare two groups (yielding a value of $df_{\text{hyp}} = 1$), using a very small sample ($N = 5$, which yields $df_{\text{err}} = 3$). Here, the study asked for the *F* and equivalent *PV* values for a traditional null hypothesis test. It is found that an *F* value of 10.13 (which would translate to $t = 3.18$) is needed to achieve statistical significance with a sample that small, and a truly large effect will be needed in the study ($PV = .772$) to get an *F* that large. In order to achieve power of .80 with a sample this small, a population effect that is even larger is necessary (i.e., to achieve power of .80 for testing the nil hypothesis with $N = 5$ and $\alpha = .05$, the population *PV* will have to be at least .858). The bottom-line message of this analysis is that researchers should not expect to reject the null with a sample that small unless the effect of their treatments or interventions is truly staggering.

This program can also be used to quickly determine the sample size needed to achieve particular levels of power, given what is known or assumed about the population value of *PV*. For example, suppose researchers want power of .80 for tests of the significance of the difference between two treatments (i.e., the plan is to carry out a *t* test) and the treatment likely accounts for about 10% of the variance in the population. Here, $df_{\text{hyp}} = 1$, so all that is needed is to try out different values for df_{error} until the point where the *PV* estimate is approximately .10. For example, if you try a value of 50 for df_{error} , they will find that the population *PV* needed to provide power of .80 is .14. This suggests that a larger sample is needed. If researchers try a value of 100 for df_{error} , they will find that the population *PV* needed to provide power of .80 is .07. This suggests that the sample size needed is somewhere between 50 and 100. A df_{error} value of 70 results in a $PV = .10$. Because the value of df_{error} in a *t* test is equal to $N - 2$, this means that a sample with $N = 72$ (or with 36 people assigned to each of the two treatments) will provide power of .80 for tests of the null hypothesis in populations where treatments account for 10% of the variance in outcomes.

EFFECT SIZE CONVENTIONS FOR DEFINING MINIMUM-EFFECT HYPOTHESES

Earlier in this chapter, it was noted that the best choice for *PV* values that defined negligible effects might depend on a number of factors, and different values might make sense in different areas of research. However, some *PV* values seem more likely than others to be chosen in defining negligible effects. In particular, many researchers will choose either 1% of the variance (or less) or 5% of the variance (or less) as their operational definition of a negligible effect, which is why these particular values are built into the One-Stop *F* Table. These specific numbers are not cast in stone (any more than the commonly used *p* values of .05 and .01), but they do represent conventions that are somewhat widely used to describe effect sizes in the behavioral and social sciences.

Cohen's books and papers on statistical power analyses (e.g., J. Cohen, 1988) have suggested a number of conventions for describing treatment effects as "small," "medium," or "large." These conventions are based on surveys of the literature, and seem to be widely accepted, at least as approximations. Table 2.2 presents conventional values for describing large, medium, and small effects, expressing these effects in terms of a number of widely used statistics.

For example, a small effect might be described as one that accounts for about 1% of the variance in outcomes, or one where the treatment mean is about one fifth of a standard deviation higher in the treatment group than in the control group, or as one where the probability that a randomly selected member of the treatment group will have a higher score than a randomly selected member of the control group is about .56.

The values in Table 2.2 are approximations and nothing more. In fact, a few minutes with a calculator shows that they are not all exactly equivalent (e.g., if you square an *r* value of .30, the result is an estimate of $PV = .09$, not $PV = .10$). Although they are not exact or completely consistent, the values in Table 2.2 are nevertheless very useful. These conventions provide a starting point for statistical power analysis, and they provide a sensible basis for comparing the results in any one study with a more general set of conventions.

The One-Stop *F* Table includes *F* statistics for testing the substantive hypotheses that the effects of treatments are either at least

TABLE 2.2

Some Conventions for Defining Effect Sizes

	PV	r	d	f^2	*Probability of a Higher Score in Treatment Group*
Small effects	.01	.10	.20	.02	.56
Medium effects	.10	.30	.50	.15	.64
Large effects	.25	.50	.80	.35	.71

Note: Cohen's $f^2 = R^2/(1-R^2) = \eta^2/(1-\eta^2) = PV/(1-PV)$, where $\eta^2 = SS_{\text{treatments}}/SS_{\text{total}}$. Data from J. Cohen (1988) and Grissom (1994).

small (i.e., 1% of the variance) or are at least small to moderate in size (i.e., 5% of the variance).

It is possible to generate critical *F* values for any other set of conventions (Appendix A shows methods of calculating the relevant noncentral *F* values for different *PV* values), but the utility of testing more and more stringent hypotheses (e.g., the hypothesis that the effect of treatments is at least large) diminishes as the effect size of interest increases. When working with strong effects, statistical hypothesis tests (especially tests of nil hypotheses) may become less and less informative. More to the point, most research in the social and behavioral sciences is likely to involve smaller rather than larger effects.

CONCLUSIONS

The statistics most widely used in the social and behavioral sciences are either interpreted in terms of, or easily translated into, the *F* statistic. The model for power analysis uses the noncentral *F* distribution to estimate the power of a wide range of statistics (cf. Patnaik, 1949). This noncentral *F* represents the distribution of outcomes that would be expected to be found in any particular study (given an effect size, df_{hyp} and df_{err}); the degree of noncentrality (λ) is a direct function of the effect size of interest. The statistical power of the study is simply the proportion of this noncentral *F* distribution that lies above whatever criterion used to define statistical significance. This model of power analysis is not tied to tests of the traditional null hypothesis (i.e., that treatments had no effect whatsoever), but rather can be easily generalized to tests of substantively meaningful hypotheses (e.g., that the treatment effect exceeds some specific value).

Analytic and tabular methods of statistical power analysis were discussed. In particular, the One-Stop *F* Table was introduced, which contains all of the information needed to: (a) test the traditional null hypothesis, (b) estimate statistical power for testing this traditional null, (c) test the substantive null hypothesis that the effect of treatments exceeds a sensible standard for defining "negligible" effects (e.g., that treatments account for no more than 1% of the variance in outcomes), and (d) test the hypothesis the substantive null hypothesis that the effect of treatments exceeds a sensible standard for defining "small to moderate" effects (i.e., that treatments account for no more than 5% of the variance in outcomes). A parallel table that was discussed lists *PV* values equivalent to the *F*s shown in the One-Stop F Table. These allow researchers to conduct analyses in terms of effect size estimates rather than in terms of their *F* equivalents. Both tables contain the same basic information, but different users might find one form or the other more convenient to use. The

One-Stop *F* Calculator program was also discussed. This program gives a simple interface for carrying out virtually all of the analyses discussed in this chapter. Finally, effect size estimates, which are widely accepted as definitions of "small," "medium," and "large" effects in the social and behavioral sciences, were discussed and these were related to statistical power analysis.

3 Using Power Analyses

Chapter 1 noted that there are two general ways that power analysis might be used. First, power analysis is an extremely useful tool for planning research. Critical decisions, such as how many subjects are needed, whether multiple observations should be obtained from each subject, and even what criterion should be used to define statistical significance can be better made by taking into account the results of a power analysis. Decisions about whether or not to pursue a specific research question might even depend on considerations of statistical power. For example, if the research idea involves a small (but theoretically meaningful) interaction effect in a complex four-way ANOVA, power analysis might show that thousands of subjects would be needed to have any reasonable chance of detecting the effect. If the resources are not available to test for such an effect, then it is certainly better to know this before the fact than to learn it after collecting the data.

Second, power analysis is a useful diagnostic tool. Tests of the traditional null hypothesis often turn out to be little more than roundabout power analyses. If a study is conducted and all of the correlations among variables, as well as all of the planned and unplanned comparisons between treatments turn out to be statistically significant, then this probably indicates that the sample was very large. If the sample is large enough, then there will be tremendous power and any effect that is literally different from zero will also be statistically different from zero. On the other hand, if none of researchers' well-conceived hypotheses are supported with significant results, they might want to conduct a power analysis before asking what is wrong with their ideas. If the power is too low, they might never reject the null hypotheses, even in cases where it is clearly and obviously wrong.

This chapter discusses the major applications of power analysis. Chapter 1 noted that because statistical power is itself a function of three parameters, the number of observations (*N*), the criterion used to define statistical significance (α), and the effect size (*ES*), it is possible to solve for any one of four values (i.e., power, *N*, *ES*, or α), given the other three. The effect size parameter may be the most problematic, because it represents a real but unknown quantity (i.e., the real effect of treatments). Before discussing practical applications of power analysis, it is useful to examine more closely the methods that might be used in estimating effect sizes.

ESTIMATING THE EFFECT SIZE

Chapter 1 noted that the exact effect size is usually unknown; if researchers knew precisely how treatment groups would differ, there would be little point in carrying out the research. There are three general methods that might be followed in estimating effect sizes in statistical power analysis. First, they might use inductive methods. If similar studies have been carried out before, they might use the results from these studies to estimate effect sizes in their own studies. Twenty years ago, this inductive method might have relied heavily on personal experience (i.e., whatever studies a particular researcher has read and remembers), but with the rapid growth of meta-analysis, it is often easy to find summaries of the results of large numbers of relevant studies (see, e.g., J. E. Hunter & Hirsh, 1984; Lipsey & Wilson, 1993), already translated into a convenient effect size metric (e.g., *d*, r^2, or more generally, *PV*).

Second, deductive methods might be used, in which existing theory or findings in related areas are used to estimate the size of an effect. For example, suppose researchers want to estimate the effect of vitamin supplements on performance in long-distance races. Suppose further that they know that the vitamin supplement has a strong and immediate effect on the efficiency with which your body uses oxygen, and efficiency in using oxygen is strongly correlated with success in such a race. It seems reasonable in this context to deduce that the vitamin supplements should have a strong influence on race outcomes.

Third, they might use widely accepted conventions about what represents a large, medium, or small effect to structure the power analysis. As noted later, analyses based on these conventions require very careful thought about what sort of effect may be realistically expected or what sort of information about statistical power is really needed.

Nevertheless, the use of these conventions can help in carrying out useful and informative power analyses, even if there is no basis for predicting with any accuracy the size of the treatment effect.

Inductive Methods

Inductive methods are best where there is a wealth of relevant data. For example, there have been hundreds and perhaps thousands of studies on the validity of cognitive ability tests in predicting performance in school and on the job (J. E. Hunter & Hirsch, 1987; J. E. Hunter & R. F. Hunter, 1984). Similarly, Lipsey and Wilson (1993) reviewed numerous meta-analyses of psychological and educational interventions. Nor is this method restricted to the behavioral and social sciences; meta-analytic procedures are being applied in areas such as cancer research (e.g., Himel, Liberati, Gelber, & Chalmers, 1986); Lipsey and Wilson (1993) cited numerous meta-analyses of research on other medical topics.

Suppose research is being conducted on the effectiveness of programs designed to help individuals quit smoking and there are sufficient resources to collect data from 250 subjects. Lipsey and Wilson (1993) cited two separate meta-analyses that suggest relatively small effects on quit rates (for physician-delivered and worksite programs, $d = .34$ and $d = .20$, based on 8 and 20 studies, respectively). This body of research provides a reasonable starting point for estimating power; a weighted mean of these two d values is .24. The F equivalent for this effect size estimate, given a study comparing quit rates in a treatment group ($n = 125$) with those in a control group ($n = 125$) is $F(1, 248) = 3.57$. The power for testing the traditional null hypothesis ($\alpha = .05$) is therefore just below .50 (the F equivalent for power of .50 from the One-Stop F Table is 3.81). If the hypothesis that treatments accounted for 1% or less of the variance in outcomes in the population is being tested, the power would be well below .50 (the F equivalent for power of .50 is 10.33; in this study, power is lower than .10).

Even though the study under consideration has a relatively large sample ($N = 250$), power is low, even for testing the traditional null hypothesis. The reason for this is that the effects of smoking cessation programs on quit rates are small ($d = .24$ means that treatments account on average for about 1.4% of the variance in quit rates). The body of research in this area provides good reason to expect small effects, and if researchers want to have adequate power for detecting these effects, they will need a much larger sample (e.g., you will need about 550 subjects will be needed to achieve power of .80 in tests of the traditional null hypothesis).

Deductive Methods

Deductive methods are the best when there is a wealth of relevant theory or models; Newtonian mechanics is probably the best example of an area in which numerous effects can be deduced on the basis of a small set of theoretical statements and principles. However, there are areas in the social and behavioral sciences where the relevant theories or models are sufficiently well developed that sound inferences could be made about effect sizes. For example, the models most commonly used to describe human cognitive abilities are hierarchical in nature, with specific abilities linked to broad ability factors, which in turn are linked to a single general cognitive ability (see Carroll, 1993, for a review of factor analytic studies). If researchers wanted to estimate the validity of a new test measuring specific abilities as a predictor of performance in school, they could use what is known about the structure of abilities to make a reasonable estimate; tests strongly related to verbal or to general cognitive ability factors are likely to show moderate to strong relations to school performance.

Suppose they used existing models of cognitive ability to estimate test validity and obtained a figure of .40 (i.e., estimated correlation between test scores and school performance). If a sample of 65 subjects is used, an expected correlation of .40 would yield an F equivalent value of $F(1, 62) = 11.80$, and the power for testing the traditional null would be greater than .80; tests of the hypothesis that tests account for more than 1% of the variance in the population would also have power of in excess of .80. Because the expected effect is relatively large, it is easy to obtain adequate power, even with a small sample.

Effect Size Conventions

As noted in the preceding chapter, there are some widely accepted conventions for defining small, medium, and large effects. For example, a *small* treatment effect has been described as one in which treatments account for approximately 1% of the variance in outcomes, the difference between treatment and control group means is about one fifth of a standard deviation, or as one in which a person randomly selected from the treatment group has a probability of .56 of having a higher score than a person randomly selected from the control group (see Table 2.2). None of these figures is sacred or exact (e.g., 2% of the variance might reasonably be described as a small effect), but the conventions described in Table 2.2 do seem to be accepted as reasonable by many researchers, and they provide a basis for doing a power analysis even when the actual treatment effect cannot be estimated by inductive or deductive methods.

When conventions are used to estimate effect sizes, it is usually best to base power analyses on small or small-to-medium effect sizes. As noted in chapter 1, a study with sufficient power to detect a small effect will have sufficient power for detecting medium and large effects as well. If researchers plan their study with the assumption that the effect is a large one, they run the considerable risk of missing meaningful effects that did not reach quite the magnitude they optimistically hoped to achieve. Thus, if the data and theory in a particular field do not provide a firm inductive or deductive basis for estimating effect sizes, they can always follow the convention and base their analysis on the assumption that the effects might very well be small. A study with sufficient power to reliably detect small effects runs little risk of making a serious Type I or Type II error, regardless of the actual size of the treatment effect.[1]

FOUR APPLICATIONS OF STATISTICAL POWER ANALYSIS

The two most common application of statistical power analysis are in: determining the power of a study, given *N*, *ES*, and a; and determining how many observations will be needed (i.e., *N*), given a desired level of power, an *ES* estimate, and an a value. Both analyses are extremely useful in planning research, and are usually so easy to do that they should be a routine part of designing a study. Power analysis may not be the only basis for determining whether to do a particular study or how many observations should be collected, but a few simple calculations are usually enough to help researchers make informed decisions in these areas. The lack of attention to power analysis (and the deplorable habit of placing too much weight on the results of small sample studies) are well documented in the research literature (J. Cohen, 1962; Haase et al., 1982; Sedlmeier & Gigerenzer, 1989), and there is no good excuse to ignore power in designing studies.

There are two other applications of power analysis that are less common, but no less informative. First, power analysis can be used to evaluate the sensitivity of studies. That is, power analysis can tell what sorts of effect sizes might be reliably detected in a study. If the effect of a treatment is expected to be small, then it is important to know whether the study will detect that effect, or whether it has sufficient sensitivity to detect only larger effects. Second, power analysis can be

[1] We noted in chapters 1 and 2, Type I errors are a concern only when there is some possibility that the null hypothesis is true, which is virtually never the case in tests of the traditional null. Type II errors are still possible if the treatment effect is extremely small, but this type of error (i.e., concluding that treatments have no effect when in fact they have a completely trivial effect) would not be regarded as very serious.

used to make rational decisions about the criteria used to define statistical significance.

Calculating Power

Chapters 1 and 2 were largely devoted to explaining the theory and procedures used in calculating the statistical power of a study, and these details are not repeated here. It is, however, useful to comment on problems or issues that arise in carrying out each of the steps of statistical power analysis.

As the examples presented in chapter 2 suggest, researchers should do the following when estimating the level of power expected in a study:

1. Estimate the size of the effect, expressed in terms of a common *ES* measure, such as *PV* or *d*.

2. Determine the degrees of freedom (i.e., df_{hyp} and df_{err}) for the statistical test to be performed, and the type of hypothesis (e.g., traditional null, minimum-effect hypothesis) to be tested.

3. Translate that effect size estimate into an *F* equivalent, given these df_{hyp} and df_{err} values.

4. Use the One-Stop *F* Table to estimate power. If the *F* equivalent is greater than the *F* needed to obtain power of .80, there should be sufficient power for most purposes, and a more precise estimate is probably not necessary. If the *F* equivalent is smaller than the *F* needed to obtain power of .50, there will not be sufficient power for most purposes, and a more precise estimate is probably not helpful. If the *F* equivalent is between these two values on the One-Stop *F* Table, the interpolation formulas in chapter 2 (Equations 9 and 10) can be used to estimate the power level of the study.

As noted earlier, there are several ways to estimate the effect size. Regardless of the method chosen, it is usually better (or at least more prudent) to underestimate than to overestimate ES values; a study with enough power to detect a small effect will also have enough power to detect a larger effect. Second, when planning a study, some preliminary estimate of the sample size (and possibly of the research) design must be made. This estimate, which may be modified if the study yields either too much or too little power, helps in determining the degrees of freedom of the *F* statistic. In most cases, power analyses are likely to lead to an increase in the sample size, but it is certainly possible that a power analysis will lead to the conclusion that adequate power can be obtained with a smaller sample than the one you had initially planned.

As emphasized in chapters 1 and 2, power depends substantially on the precise hypothesis being tested. It is easier to obtain high levels of power for tests of the traditional null hypothesis than for tests of a minimum-effect hypothesis, at least in part because the traditional null is so easily rejected. Tests of minimum-effect hypotheses, although more difficult, are also likely to be more informative.

Finally, a decision must be made about the significance criterion. Later, this chapter discusses in detail the issues involved in making this choice. Here, it is sufficient to note that the choice is often practically limited to the conventional values of .05 versus 01. If any other value is used to define statistical significance (e.g., .02 might be a perfectly reasonable choice in some settings), researchers will have to fight convention and defend their choice to a potentially hostile set of reviewers, readers, and editors (Labovitz, 1968).

Determining Sample Sizes

Rather than calculating the level of power a particular study had or will have, it is often useful to determine the sample or research design needed to achieve specific levels of power. To do this, it is necessary to decide how much power is desired. As noted in chapter 2, it is hard to justify a study design that yields power less than .50; when power is less than .50, the study is more likely to lead to an incorrect conclusion (i.e., it will not reject H_0, even though it is virtually certain this hypothesis is wrong) than to a correct one. Power substantially above .80 might be desirable, but it is often prohibitively difficult to obtain; in most analyses, the desirable level of power is likely to be .80.

Once the desirable level of power is selected, determining sample sizes follows the same general pattern as the determination of power itself. That is, to take into account the research design, the estimated effect size, the nature of the hypothesis being tested, and the significance criterion being used. Appendix D can be used to determine sample sizes needed to detect a wide range of effects.

The rows of Appendix D correspond to effect sizes, described in terms of either the standardized mean difference (*d*) or the proportion of variance explained (*PV*), which represent the most common effect size estimates in the literature. Table 3.1 provides formulas for translating a number of other statistics, including *F*, into *d* and/or *PV*.

The values in Appendix D represent degrees of freedom (df_{err}) rather than exact sample sizes (*N*). The reason for presenting a table of df_{err} values rather than presenting a table of sample sizes is that *N* is a function of both df_{hyp} and df_{err}. In many applications, $N = df_{\text{hyp}} + df_{\text{err}} +$

TABLE 3.1

Translating Common Statistics Into Standardized Mean Difference (*d*) or Percentage of Variance (*PV*) Values

Standardized Mean Difference (d)
$d = \frac{2r}{\sqrt{1 - r^2}}$
$d = \frac{2t}{\sqrt{df_{err}}}$

Percentage of Variance (PV)
$PV = \frac{t^2}{t^2 + df_{err}}$
$PV = \frac{df_{hyp}F}{df_{hyp}F + df_{err}}$
$PV = \frac{d^2}{d^2 + 4}$

1, but in complex multifactor designs (e.g., studies using factorial ANOVA), the total sample size depends on the number of levels of all design factors, and without knowing the research design in advance, it is impossible to put together an accurate *N* needed table. In most cases, however, the sample size needed to achieve power of .80 will be very close to the df_{err} value shown in Appendix D. Chapter 5 deals with applications of power analysis in multifactor studies.

Appendix D presents df_{err} needed when testing the traditional null hypothesis. Appendix E presents the df_{err} needed when testing the hypothesis that treatments account for 1% or less of the population variance in outcomes. So, larger samples are needed when testing this minimum-effect hypothesis than when testing the traditional (but sometimes trivial) hypothesis that treatments have no effect whatsoever.

To illustrate the use of Appendix D and Appendix E, consider a study comparing the effectiveness of four diets. Suppose a small to moderate effect (e.g., the choice of diets is expected to account for about 5% of the variance in weight loss) is expected. To achieve power of .80 in testing the traditional null hypothesis (with α = .05), look down the column of Appendix D that corresponds to $df_{\text{hyp}} = 3$ (i.e., with four diets, there are three degrees of freedom) and find that the

df_{err} needed for $PV = .05$ would be about 209. A sample of about 214 subjects (i.e., $N = 209 + 3 + 1$) would be necessary to achieve this level of power, or abut 54 per group.

Appendix E shows the df_{err} needed to achieve power of .80 in testing the hypothesis that treatments account for 1% or less of the variance in outcomes. To achieve a power of .80 in testing this minimum-effect hypothesis, a sample of about 409 (i.e., $N = 405 + 3 + 1$), or about 100 per group, would be required. This sample is almost twice as large as the one needed to reject the traditional null hypothesis that treatments have no effect. However, if a study allows researchers to reject the hypothesis that treatments have a negligibly small effect (e.g., they account for less than 1% of the variance in outcomes), they will not only know that treatments have some effect, but they will also have a formal test of the hypothesis that the effect is large enough to warrant attention.

Determining the Sensitivity of Studies

It is often useful to know what sort of effect could be reasonably detected in a particular study. If a study can reliably detect only a large effect (especially in a context where small effects are expected to actually occur), it might be better to postpone that study until the resources needed to obtain adequate power are available. The process of determining the effect size that can be detected—given particular values for N and a, together with a desired level of power—again closely parallels the procedures described earlier. In fact, Appendices D and E, which specify the df_{err} needed to achieve power of .80 in testing both traditional and minimum-effect hypotheses, are also quite useful for determining the type of effect that could be reliably detected in a given study.

Consider a comparison of two methods of mathematics instruction. There are 140 students available for testing (70 will be assigned to each method), and the decision is to use an a level of .05 in testing hypotheses. Here, $df_{\text{hyp}} = 1$ (i.e., if two treatments are compared, there is one degree of freedom for this comparison) and $df_{\text{err}} = 138$ (i.e., $df_{\text{err}} = N - df_{\text{hyp}} - 1$). Looking down the $df_{\text{hyp}} = 1$ column of Appendix D, it is found that 138 falls somewhere between the df_{err} needed to detect an the effect of treatments when $PV = .05$ and $PV = .06$ (or between $d = .46$ and $d = .51$). In other words, with 140 students, there would be a power of .80 to detect a small to moderate effect, but there would not be this level of power for detecting a truly small (but perhaps important) effect.

Take another example. Suppose there are four cancer treatments and 44 patients. Here, $df_{\text{hyp}} = 3$ and $df_{\text{err}} = 40$. Appendix D indicates that there is a power of .80 to detect effects that are quite large (i.e., PV

= .24); if the true effects of treatments are small or even moderate, the power to detect these effects will drop substantially.

Both of the aforementioned examples are based on tests of the traditional null hypothesis. To test the hypothesis that treatments account for 1% or less of the variance in outcomes (which corresponds to d values of .20 or lower), the sample of 140 students will give the power of .80 for detecting differences this large between the two treatments only if the true effect is relatively large (i.e., $PV = .09$ or $d = .63$; see Appendix D). This level of power will exist for detecting nontrivial differences among the four cancer treatments (with $N = 44$) only if the true differences between treatments are truly staggering (i.e., $PV = .26$; see Appendix E).

Determining Appropriate Decision Criteria

As noted earlier, the choice of criteria for defining statistical significance is often practically limited to the conventional values of .05 versus .01 (occasionally, social scientists use .001 or .10 as significance levels, but these are rare exceptions). The choice of any other value (e.g., .06) is likely to be met with some resistance, and the battle is probably not worth the effort. Because the choice among significance levels is constrained by convention, the steps involved in making this choice do not exactly parallel the processes laid out in the three preceding sections of this chapter. Rather than describing specific steps in choosing between .05 and .01 as alpha levels, consider the range of issues that are likely to be involved in making this choice. This discussion leads to the conclusion that researchers should never use the .01 level when testing traditional null hypotheses, nor should they usually use other procedures designed to guard against Type I errors in testing this hypothesis. As shown later, choice of the .01 significance criterion leads to a substantial reduction in statistical power, with virtually no meaningful gain in terms of protection against Type I errors. The same is true of most procedures designed to reduce Type I errors (see Zwick & Marascuilo, 1984, for a review of procedures used in testing multiple contrasts).

Balancing Risks in Choosing Significance Levels Chapter 1 revealed that when testing the traditional null hypothesis, two types of errors are possible. Researchers who reject the null hypothesis when in fact it is true make a Type I error (a is the probability of making this error if H_0 is in fact true). The practical effect of a Type I error is that researchers could come to believe that treatments have some effect when in fact they have no effect whatsoever. Researchers who fail to reject the null hypothesis when it is false make a Type II error (β is the probability of making this error when *Ho* is in fact false, and power = 1 − β). The practical effect of making a Type II error is that researchers might give up on treatments that in fact have some effect.

The most common strategy for reducing Type I errors is to make it difficult to reject the null hypothesis (e.g., by using .01 rather than .05 as a criterion for significance). Unfortunately, this strategy also substantially reduces the power of tests. For example, suppose researchers are comparing two treatments (with 200 people assigned to each treatment) and they expect a small effect (i.e., d = .20). Using .05 as a significance criterion, their power would be .64; if α = .01, power drops to .37 (J. Cohen, 1988). This trade-off between Type I error protection and power suggests that in deciding which significance level to use, they must balance the risk and consequences of a Type I error with the risk and consequences of a Type II error. Nagel and Neff (1977) discussed a decision-theoretic strategy for choosing an alpha level that provides an optimum balance between the two errors.

Cascio and Zedeck (1983) suggested that Equation 12 can be used to estimate the apparent relative seriousness (ARS) of Type I versus Type II errors in statistical significance tests:

$$ARS = \frac{p(H_1)\beta}{\left[1 - p(H_1)\right]\alpha} \qquad [12]$$

where:

$p(H_1)$ = probability that H_0 is false

According to Equation 12, if the probability that treatments have some effect is .7, α is .05 and the power is .80, a mistaken rejection of the null hypothesis (i.e., a Type I error) is treated as if it was 9.33 times as serious [i.e., (.7 * .2)/(.3 * .05) = 9.33] as the failure to reject the null when it is wrong (i.e., a Type II error). In contrast, setting a at .10 leads to a ratio of 4.66, or to the conclusion that Type I errors are treated as if they are 4.66 times as serious as a Type II error (See also Lipsey, 1990).

The first advantage of Equation 12 is that it makes explicit values and preferences that are usually not well understood, either by researchers or by the consumers of social science research. In the scenario described, an alpha level of .05 makes sense only if researchers think that Type I errors are over nine times as serious as Type II errors. If they believe that Type I errors are only four or five times as serious as Type II errors, then they should set their significance level at .10, not at .05.

The second advantage of Equation 12 is that it explicitly involves the probability that the null hypothesis is true. As noted in several other contexts, the traditional null hypothesis is virtually never true, which means that both null hypothesis testing and efforts to reduce Type I errors are sometime pointless (Murphy, 1990). If the null hypothesis is by definition false, then it is not possible to make a Type I error. Thus, the only circumstance in which stringent significance criteria should be used, testing procedures that minimize type I errors should be adopted, and so on. are those in which the null hypothesis might actually be true. This virtually never happens when testing the traditional null hypothesis, but it might occur when testing a minimum-effect hypothesis.

Should You Ever Worry About Type I Errors? In testing the traditional null hypothesis, it is often virtually impossible to make a Type I error. If *Ho* is known to be false, then there simply is no way to make a Type I error, and the only concern should be maximizing power (and therefore minimizing Type II errors). In contrast, there are good reasons for concern over Type I errors when testing a minimum-effect hypothesis. It is virtually impossible to devise serious treatments or interventions that have no effect whatsoever, but there are many treatments, interventions, tests, and so forth. that have only negligible effects. If the hypothesis to be tested is that the effect of treatments falls under some sensible minimum, then it is quite possible that the null will be true, and something will be learned by testing it.

As Equation 12 suggests, it is impossible to sensibly evaluate the relative emphasis given to Type I versus Type II errors unless the probability that the null hypothesis is in fact true can be estimated. Table 3.2 shows the relative seriousness with which Type I versus Type II errors are treated, as a function of both the alpha rate and the probability that the null hypothesis is true, in studies where power = .80. For example, if there is a 5% chance that treatments really do have a negligible effect (i.e., the probability that the minimum-effect null hypothesis is true is .05), the decision to use an alpha level of .01 makes sense only if it is believed that Type I errors are 380 times as serious as Type II errors. As indicated in Table 3.2, as a increases (i.e., as it becomes easier to reject H_0), the relative seriousness attached to Type I versus Type II errors goes down. So, if the belief is that Type II errors are at all serious relative to Type I errors, the message of Table 3.2 is clear; Use a more lenient criterion for determining statistical significance.

TABLE 3.2

Relative Seriousness of Type I Versus Type II Errors as Function of Alpha and Probability that the Null Hypothesis is True (Power = .80)

Probability that HO is True	*Alpha (α)* *.01*	 *.05*	 *.10*
.50	20.00	4.00	2.00
.30	46.66	9.33	4.66
.10	180.00	36.00	18.00
.05	380.00	76.00	38.00

It is easy to rearrange the terms of Equation 12 to compute the alpha level that *should* be used to reach an appropriate balance between Type I and Type II errors, a balance described as the *desired relative seriousness* (DRS).[2] For example, if researchers decided that they wanted to treat Type I errors as twice as serious as Type II errors, the desired relative seriousness is 2.00; if the probability that the null hypothesis was true was .30, power was .80, and they wanted to treat Type I errors as if they were twice as serious as Type II errors, researchers should use an alpha level of .23. The desired alpha level (i.e., level that will yield the appropriate balance between Type I and Type II errors) can be obtained from Equation 13:

$$\alpha_{desired} = \left[\frac{p(H_1)\beta}{1-p(H_1)}\right]\left(\frac{1}{DRS}\right) \quad [13]$$

where:

α_{desired} = alpha level that will yield the desired relative seriousness of Type I and Type II errors
DRS = desired relative seriousness of Type I and Type II errors

[2] Keep in mind, however, that using nonconventional alpha levels will often require vigorous defense of this choice, even in contexts where the "conventional" choice (e.g., .05, .01) makes no sense whatsoever.

To repeat a critical point, be concerned with Type I errors if and only if there is a realistic possibility that such an error can occur. One of the distinctions between the traditional null and minimum-effect tests is that there is a realistic possibility that the minimum-effect null hypothesis will be true, which suggests that serious decisions must be made about the appropriate balance between Type I and Type II errors. Equations 12 and 13 can help in making those decisions.

CONCLUSIONS

Statistical power analysis can be used in planning future studies (e.g., determining how many subjects are needed) and/or in diagnosing studies that have already been carried out (e.g., calculation of power levels helps researchers sensibly interpret significance tests). All applications of statistical power analysis require at least an estimate of the true effect size; estimates can be obtained form literature reviews or from relevant theory. Even where little is known about the true effects of treatments, effect size conventions can be used to structure power analyses.

It is possible to use the methods described here to determine the probability that a study will correctly reject the null hypothesis, determine the number of subjects or observations needed to achieve adequate power, determine the sort of effects that can be reliably detected in a particular study, or make informed choices about the criteria that define statistical significance. As noted throughout this book, these applications of power analysis are not in any way limited to tests of the traditional null hypothesis, but rather can be easily adapted to more interesting and informative tests of the hypothesis that treatment effects exceed some minimum level.

The distinction between traditional and minimum-effect hypotheses is especially important when making decisions about criteria for defining a significant outcome. Simply put, when testing the traditional null hypothesis, there is rarely any justification for choosing a stringent alpha level (e.g., .01 rather than .05). Procedures designed to protect against Type I errors (e.g., Bonferroni corrections) usually reduce power, and should only be applied if there is some realistic possibility that a Type I error can be made. This is unlikely when testing the traditional null hypothesis; this hypothesis is usually known to be incorrect, virtually by definition. Type I errors are a very real concern in tests of minimum-effect hypotheses, and the equations presented here allow researchers to rationally assess the relative emphasis placed on maximizing statistical power versus avoiding errors of this sort.

4

Multi-Factor ANOVA and Repeated Measures Studies

Experiments and quasi-experimental studies in the behavioral and social sciences often involve several independent variables that are manipulated or studied jointly. For example, an educational psychologist might be interested in the effects of both the amount of instruction (e.g., 2, 3, or 4 days a week) and the method of instruction (e.g., traditional lecture vs. hands-on learning) on the achievement of students. A factorial experiment is one in which the individual and joint effects of both of these independent variables can be studied. So, for example, an investigator might randomly assign each of 240 students to one of the conditions illustrated in Fig. 4.1.

	2 days	3 days	4 days
Lecture			
Hands-on			

FIG. 4.1. A factorial experiment.

If students are randomly assigned to these six conditions, the result should be 40 students in each cell of Fig. 4.1 (in this sort of study, it is common to use n to designate the number of people in each condition and N to designate the number of participants).

Sometimes, it is not possible to assign students at random. For example, it might be administratively impossible to move students around at random, but it might be possible to assign classes or groups of students to each cell, creating a quasi-experimental study. The distinction between a true experiment and a quasi-experimental study is that experiments involve random assignment of each individual to conditions, whereas quasi-experimental studies do not allow every subject to be randomly assigned to conditions. The methods used to analyze data from experiments and quasi-experiments are identical, but true experiments permit stronger inferences about causation.

THE FACTORIAL ANALYSIS OF VARIANCE

The analysis of variance (ANOVA) is a statistical technique that involves asking the question, "Why do some people get high scores and others get low scores on the dependent variable?" (or, "why do scores vary?"). In its most general form, in a study where subjects are assigned to different treatments and their scores on some common dependent variable are obtained, ANOVA answers this question by breaking down the variability of scores into two components, as illustrated in Equation 14:

$$\text{Variability in scores} = \text{Variability due to differences in treatment means} + \text{Variability not due to differences in treatment means} \qquad [14]$$

In a study where there is only one independent variable (e.g., where the independent variable is the number of days of instruction), Equation 14 corresponds to:

$$\text{Variability in scores} = \text{Variability due to treatments} + \text{Variability due to error} \qquad [15]$$

In this sort of study, the mean square between (MS_B) provides an estimate of the variability associated with treatments, the mean square within (MS_W) provides an estimate of the variability in scores among subjects who receive the same treatment (i.e., variability that is not due to the treatments). Thus, MS_W represents variability due to error (i.e., the variability that would be expected if all the differences among scores were merely random, chance fluctuations). The statistic $F = MS_B/MS_W$ tells whether the variability due to treatments is large relative to the variability in scores due to error (i.e., whether the variability due to treatments is larger than what would be expected if everything was due to chance). If the variablity due to treatments represents more than chance fluctuations, it may be concluded that the treatments had some effect. In a one-way ANOVA, all of the variability in scores is broken down into these two components (between vs. within), and as a result, it is easy to use the *F* statistic to estimate the effect size, *PV.*

As noted in chapter 2, in a one-way ANOVA, $PV = df_{hyp} * F/[(df_{hyp} * F) + df_{err})]$. In factorial ANOVA, the questions asked and the breakdown of the variability in scores becomes more complex, and as a result, the estimation of *PV* and of statistical power also becomes more complex. This chapter generalizes the formula used for estimating *PV* from the *F* values obtained in a factorial ANOVA study.

Factorial experiments allow researchers to ask several questions simultaneously. For example, the factorial study illustrated in Fig. 4.1 allows them to ask three questions: (a) What is the effect of the number of days of instruction, (b) what is the effect of the style of teaching, and (c) do these two variables interact (i.e., does the effect of the amount of instruction depend on the method of instruction used?)? Different research designs might allow researchers to ask a wider range of questions (e.g., a later section of this chapter discusses repeated measures designs, which allow them to examine systematic subject effects and subject by treatment interactions), or might provide a more narrowly focused examination of the data (e.g., designs with nested factors can be more efficient, but may not provide an opportunity to ask all of the questions that are possible in a fully crossed factorial design). In multi actor ANOVA studies, it is common to have different levels of power for each of hypotheses tested.

Different Questions Imply Different Levels of Power

In the design illustrated in Fig. 4.1, there are three distinct hypotheses that can be tested via ANOVA, two main effect hypotheses (i.e., the hypothesis that the type of instruction and the amount of instruction influence learning) and one interaction hypothesis (i.e., the hypothesis that the effects of the type of instruction depends on the amount of instruction). All other things being equal, they will have more power for testing the hypothesis that the type of instruction matters than for testing the hypothesis that the amount of instruction matters. Researchers will have more power for testing either of these main-effect hypotheses than for testing the hypothesis that the amount and type of instruction interact. To understand why this is true, it is necessary to think concretely about the type of question ANOVA is designed to answer.

In some ways, the name "analysis of variance" is an unfortunate one, because it can lead people to forget what ANOVA was designed to do. The analysis itself focuses on identifying sources of variability or variance (hence the name ANOVA), but the reason for doing the analysis is generally to compare means. For example, when asking whether the main effect for type of instruction is large, this is really asking whether the mean score of people who received lecture training is substantially different from the mean score of people who received hands-on instruction. Similarly, when asking whether the main effect for amount of instruction is large, the real question concerns whether the scores of people who received more instruction are substantially different from the scores of people who received less instruction. Interaction hypotheses in the design illustrated in Fig. 4.1 involve comparisons of individual cell means.

A general principle running through all of the power analyses discussed in chapters 1–3 is that there is more power when hypotheses are tested in large samples than when similar hypotheses are tested in smaller samples. That is, power increases with *N*. In ANOVA, the level of power for comparing means (i.e., for testing hypotheses about main effects and interactions) depends largely on the number of observations that goes into calculating each mean. Suppose, for example, that 240 subjects show up for the study illustrated in Fig. 4.1. Tests of the main effect for type of instruction will compare the mean score of the 120 people who receive lectures with the mean score of the 120 people who receive hands-on instruction. Tests of the main effect of amount of instruction will compare three means, each based on 80 subjects (i.e., 80 people received 2 days of instruction, 80 received 3 days, and 80 received 4 days). Tests of the interaction effect will involve comparing cell means, each of which is based on 40 people (e.g., 40 subjects received 2 days of lecture, another 40 received 3 days of hands-on instruction etc.). All other things being equal, power is higher for statistical tests that involve comparing means from samples of 120 people (here, tests of the main effect of instruction type) than for tests that involve comparisons samples of 80 people (here, tests of the main effect of amount of instruction). Statistical tests that involve comparisons among means obtained from samples of 40 people (here, tests of the interaction) will often have even lower levels of power.

In ANOVA, it is common to distinguish between the number of observations in each cell of a design (designated by the lowercase *n*) and the total number of observations in a study (designated by the uppercase N). In the aforementioned study, the interaction means are based on $n = 40$ subjects, and the main effect means for the type of instruction and amount of instruction main effects are based on samples of $3n$ and $2n$ subjects (i.e., 120 and 80), respectively. If all other things were equal, the larger the sample, the more power would be expected. Unfortunately, all things are rarely if ever equal, and it is difficult to state as a general principle that tests of main effects will always be more powerful than tests of interactions, or that main effect tests that involve comparing fewer means (and therefore fewer df_{hyp}) will be more powerful than main effect tests that involve comparing more means (and therefore more df_{hyp}). The reason for this is that the effect sizes for each of the separate questions pursued in ANOVA can vary. For example, suppose that there is a very large interaction effect, a moderately large main effect for the amount of instruction, and a very small main effect for the type of instruction. This might lead to the highest level of power for tests of the interaction and to the lowest level of power for tests of the type of instruction main effect.

Estimating Power in Multifactor ANOVA

The process of estimating the power of tests of main effects and interactions in ANOVA is a straightforward extension of the processes described in chapters 1 and 2. In particular, power depends on sample size, effect size, and the decision criterion. For example, assume that researchers expected a small main effect ($PV = .01$) for the type of instruction, a moderately large main effect ($PV = .10$) for the amount of instruction, and an equally large effect ($PV = .10$) for the interaction. If $N = 240$ (i.e., there are 240 subjects randomly assigned to the six cells in this design), they would expect power to equal approximately .30 for the type of instruction main effect (assuming a = .05). They would expect power levels of approximately .99 for the amount of instruction main effect and the type by amount interaction. In other words, power levels would be quite low for testing some effects and quite high for others in this study.

Estimating *PV From F* in a Multifactor ANOVA

In a one-way ANOVA, all of the variability in scores is broken down into two components, MS_B and MS_W. Once you know the value of F, *it is easy to determine the value of PV*, which reflects the proportion of variance due to differences between treatments [as noted earlier, in a one-way ANOVA, $PV = (df_{hyp} * F)/[(df_{hyp} * F) + df_{err})]$. In a multifactor ANOVA, F does not, by itself, give enough information to allow for calculation of *PV*. For example, in the study illustrated in Fig. 4.1, the *F* for the type of instruction main effect tells you how large this effect is relative to MS_{error}, but it does not tell how much of the total variance is accounted for by differences in types of instruction. In a multifactor ANOVA, the *PV* for one main effect depends in part on the size of the other main effects and interactions.

It is possible to solve for the percentage of variance explained by each of the effects in the ANOVA model. In most cases (exceptions are described later), all that is required is the *F* and the degrees of freedom for each effect in the model. Returning to the example, suppose the following is found:

Type of Instruction: F (1, 234) = 3.64
Amount of Instruction: F (2, 234) = 15.19
Type X Amount Interaction: F (2, 234) = 16.41

These three sets of *F* values and degrees of freedom provide all of the information needed to determine the *PV* for each of these three effects, as well as the *PV* for the error term.

To see how this information can be used to calculate PV, we start by noting that:

$$F = MS_{\text{hyp}} / MS_{\text{error}} \qquad [16]$$

And

$$F * df_{\text{hyp}} = SS_{\text{hyp}} / MS_{\text{error}} \qquad [17]$$

In a design where there are two main effects (label these *A* and *B*) and one interaction (labeled *AB*), it is easy to show that:

$$PV_{\text{error}} = df_{\text{error}} / [df_{\text{error}} + (F_{\text{A}} * df_{\text{A}}) + (F_{\text{B}} * df_{\text{B}}) + (F_{\text{AB}} * df_{\text{AB}})] \qquad [18]$$

If the percentage of variance explained by error is known, it follows that the rest of the variance is explained by the model as a whole (i.e., by the combined effects of the *A* effect, the *B* effect, and the *AB* interaction).

$$PV_{\text{model}} = 1 - PV_{\text{error}} \qquad [19]$$

Finally, once the combined effects of *A*, *B*, and *AB* are all known, all that remains is to determine the percentage of variance explained by each. These values are given by:

$$PV \text{ for } A = PV_{\text{model}} * [(F_A * df_A) / ((F_A * df_A) + (F_B * df_B) + (F_{AB} * df_{AB}))] \qquad [20]$$

$$PV \text{ for } B = PV_{\text{model}} * [(F_B * df_B) / ((F_A * df_A) + (F_B * df_B) + (F_{AB} * df_{AB}))] \qquad [21]$$

$$PV \text{ for } AB = PV_{\text{model}} * [(F_{AB} * df_{AB}) / ((F_A * df_A) + (F_B * df_B) + (F_{AB} * df_{AB}))] \qquad [22]$$

Applying Equations 18 through 22 to the *F* and *df* values in the example, it may be concluded that: (a) error accounted for 77.8% of the variance (Equation 18); (b) the *A* main effect, the *B* main effect, and the *AB* interaction, combined, accounted for 22.2% of the variance (Equation 19); (3) the *A* main effect accounted for 1% of the variance (Equation 20); (d) the *B* main effect accounted for 10% of the variance (Equation 21); and (e) the *AB* interaction accounted for 11% of the variance (Equation 22).

It is easy to extend Equations 18 through 22 to designs with three, four, or more factors. There are only two real limitations to these equations. First, they can only be used with fully crossed factorial designs, where an equal number of subjects is assigned to every possible combination of levels of *A*, *B*, and so on. Thus, equations of this sort will not allow for easy computation of *PV* from *F* in nested designs or incomplete designs. Corresponding calculations are possible, but laborious. Second, and more important, these equations apply only to fixed-effect models, in which all significance tests involve comparing the *MS* for each effect in the model to MS_{error}. Models that include random effects require more complex significance tests, and it is not possible to develop a simple set of formulas that cover all possible combinations of fixed and random effects. Luckily, fixed-effect models are much more common in the behavioral and social sciences than random-effects or mixed models (e.g., the default method for virtually all statistical analysis packages is to use fixed-effects models for structuring significance tests), especially in designs that do not include repeated measures.

REPEATED-MEASURES DESIGNS

Several years ago, a colleague interviewed for a job at a university where the animal learning researchers had decided to break the mold. Instead of the usual lab with rats, pigeons, and other small animals, they had a basement lab where alligators learned to make their way through a flooded maze. We are not sure how they dealt with issues of statistical power, but it is a good bet that they did not get more alligators every time a power analysis told them that *N* was too small.

There are many research areas where it is virtually impossible to gather large numbers of subjects. Sleep research often involves multiple nights, with large numbers of observations each night. Vision research often involves hundreds of trials per subject. Although the advice that researchers should aim for large *N*s applies even in these areas, it is a fact of life that research designs that involve large numbers of subjects are sometimes impossible to implement. Repeated measures designs allow researchers to obtain a lot of information from a relatively small number of subjects by gathering several pieces of data from each subject.

Most discussions of statistical power start with the assumption that subjects are randomly assigned to treatments, and one piece of data is taken from each subject. One of the hallmarks of this between-subject design is that the observations obtained are statistically independent. That is, the score achieved by Subject 2 in a study is assumed to be independent of the score achieved by Subject 1. This design has many attractive features, but it is not always the best way to design studies. In many research areas, the scientific question itself or the practical constraints involved in running subjects require that researchers obtain several observations from each subject; these are often referred to as repeated measures designs.

Power analysis suggests that the larger the *N*, the more power there will be for statistical tests. In a between-subjects study, there is no ambiguity, because *N* refers to the number of people. In a study involving repeated measures, *N* refers to the number of observations, and different ways of producing the same number of observations will have different implications for statistical power. For example, suppose the current study involves 10 subjects, each of whom provides five pieces of data. There are 50 pieces of data, but this is not statistically equivalent to a study in which there are 50 subjects, each of whom provides a single piece of data. Similarly, doubling the number of observations might have different effects, depending on the way it is done. Ten people, each doing 10 things, will yield different levels of power than 20 people each doing 5 things, or 100 people, each doing 1 thing.

Table 4.1 illustrates a key difference between repeated measures studies and studies that use between-subject designs (in which each subject provides one piece of data). Suppose a study uses a .05 significance criterion and a small to moderate effect is expected (i.e., the difference between the most effective treatment and the least effective treatment corresponds to $d = .50$), and relatively small correlations between repeated measures (i.e., r values of .20 or lower). The table shows the number of subjects that would be needed and a repeated measures study, and contrasts that to the number of subjects needed for a equivalent level of power in a between-subjects study (Maxwell & Delaney, 1990).

A comparison of these two designs suggests many interesting things. First, the total number of observations is held constant, then researchers will get more power from repeated measures designs. With two treatments, you need only 53 subjects are needed (yielding 106 observations in a repeated measures design) to attain the same level of power as is achieved with 128 subjects in a between-subjects design. If there are four treatments, the result will be the same level of power with a sample of 74 subjects, each receiving all four treatments

TABLE 4.1

N Required to Achieve Power = .80 With Maximum d = .50 and r = .20

Between-Subject		*Repeated-Measures*	
Treatments	*Number of Subjects*	*Subjects*	*Observations*
2	128	53	106
3	237	65	195
4	356	74	296
5	485	82	410
6	624	88	528

(yielding 296 observations) as would be achieved in a between-subjects study involving 356 subjects.

Table 4.2 shows a similar set of calculations, based on the assumption that the correlation between repeated measures is larger (i.e., *r* values of at least .40). It is evident that the effect of using a repeated measures design is substantially larger if the correlations among the repeated measures are relatively high than when these correlations are low (if the correlations between repeated measures are near zero, both designs will require about the same number of observations to achieve equal levels of power). In Table 4.1, if there are four treatments that yield scores that are not highly correlated (at most, $r = .20$), you would need 74 subjects (yielding 296 total observations) would be needed to achieve the same level of power as would be achieved with 356 observations in a between-subjects study. In Table 4.2, where treatments yield scores that are more highly correlated (*r* is at least .40), 57 subjects (yielding 228 observations) would achieve the same level of power as a between-subjects study with $N = 356$.

Why Does This Work?

There are several ways to explain why repeated measures studies are more powerful than similar between-subjects studies. First, as a comparison of Tables 4.1 and 4.2 might suggest, the correlation between repeated measures makes a big difference. The analysis of variance involves breaking down the variability in scores into variance due to the treatments and variance due to error, and if observations are con-

TABLE 4.2

N Required to Achieve Power = .80 with $d_{max} = .50$ and $r = .40$

Between-Subject		*Repeated Measures*	
Treatments	*Number of Subjects*	*Subjects*	*Observations*
2	128	40	80
3	237	50	150
4	356	57	228
5	485	63	315
6	624	68	408

sistent over several treatments (which is what a high correlation between repeated measures implies), there is simply less error variance. All other things being equal, there is less random variability in a set of repeated measures than in a set of independent measures, each obtained from a different subject. The principle here is similar to the central idea in traditional psychometric theory (i.e., that by grouping together several intercorrelated observations, it is possible to get a highly reliable measure of where each subject stands on the dependent variable). Repeated measures designs allow researchers to take advantage of the fact that the mean of several correlated observations is a more reliable piece of information about a subject than a single observation, which is all they would have in a between-subjects design.

Second, repeated measures designs allow for the identification and removed of sources of variance in scores that are treated as error in a between-subject design. In particular, repeated measures designs allow for estimation and removal of systematic subject effects that cannot be estimated or controlled in typical between-subject designs. Consider the following experiment.

Subjects are asked to complete a complex psychomotor task under different sorts of distraction. There are three levels of noise (loud, very loud, painfully loud) and two levels of temperature variation (very cold, very hot). You have enough time and money to collect 60 observations. Table 4.3 shows the effects that can be estimated (along with their degrees of freedom) in a between-subjects study and in a within-subjects study in which each subject participates in all six conditions.

TABLE 4.3

Sources of Variance in Between versus Within-Subjects Designs

Between	*df*	*Within*	*df*
Noise (*N*)	2	Noise	2
Temperature (*T*)	1	Temperature (*T*)	1
N |×| *T*	2	*N* |×| *T*	2
Error	54	Subjects (*S*)	9
		N |×| *S*	18
		T |×| *S*	9
		Error	18

The repeated measures design allows researchers to find out whether there are systematic subject effects (some subjects might be better at this sort of task than others), subject by noise interactions (some subjects may be more distracted by noise than others), and subject by temperature interactions (some subjects may be more distracted by hot or cold than others). These subject effects can all be estimated and removed from the residual error term. In contrast, between-subject designs lump all of these effects into "Error." If there are meaningful subject effects and subject by treatment interactions, repeated measures designs will allow researchers to remove them from error, and between-subjects designs will treat these systematic effects as part of the overall error term. Anything they can do to reduce error is likely to increase statistical power, and one reason for the power of repeated measure designs is that they allow researchers to isolate these subject effects. Finally, there is a more general statistical explanation for the power of repeated measures designs. Chapter 2 discussed the noncentral *F* distribution, and noted that as the noncentrality parameter gets larger, the mean of the *F* distribution increases and the variance grows larger as well. One thing that affects the noncentrality parameter is the effect size; it is this relation that allowed for the creation *F* tables for minimum-effect *F* tests. Thus, the larger the effect of the interventions and treatments, the stronger the upward shift in the distribution of *F* values expected to be found in the study.

Vonseh and Schork (1986) noted that in repeated measure studies, the value of the noncentrality parameter depends on both the effect size and the correlations among repeated measures. The precise effects of the correlations among repeated measures on the distribution of *F* are complex and nonlinear, but in general the higher the correlation among measures, the more noncentral the distribution of *F* becomes.

Repeated measures studies tend to yield larger values of *F*, and therefore more power.

Why Doesn't Everyone Use Repeated Measures Designs?

Repeated measures designs are more powerful, more efficient, and easier to implement than between-subjects designs. For example, suppose a power analysis reveals that 400 subjects are needed to obtain a reasonable level of power. Researchers might go out and recruit 400 individuals, getting a single score from each one (a between-subjects design). Alternately, they might need to recruit only 10 subjects, and get 40 observations from each (in fact, this will yield more power than a between-subjects design with $N = 400$). Given the practical and statistical advantages of repeated measured designs, they might wonder why anyone uses between-subjects designs.

Although repeated measures designs are attractive in many ways, they often turn out to be inappropriate for many research topics. Consider, for example, a study examining the effectiveness of five different methods of instruction for teaching second-grade students basic mathematics. A repeated measures design that exposes all students to all five methods will run into a serious problem, often labeled a carry-over effect. That is, whatever the students learn with the first method to which they are exposed will affect their learning and performance on subsequent methods. By about the fourth time students have gone over the material (using each of four methods), they probably will have learned it, and will certainly be getting pretty sick of it. The first four trials will almost certainly affect outcomes on the fifth trial.

In general, many research questions make a repeated measures design difficult to implement. Studies of learning, or interventions designed to change attitudes, beliefs, or the physical or mental state of subjects might all be difficult to carry out in a repeated measures framework. If the interventions under study change subjects, it might be difficult to interpret data when several different interventions are implemented, one after the other. Similarly, some research paradigms involve tasks that are complex or time consuming, and even if it is theoretically feasible to repeat the task under a variety of conditions, there may be serious practical barriers to obtaining multiple observations from each subject. Some studies allow for mixed designs, in which some of the factors studied represent repeated measures and others represent between-subject factors, and these designs often represent the best compromise between the power and efficiency of repeated measures designs and the independence of observations that characterizes between-subject designs.

One of the apparent strengths of repeated measures designs is also a potential weakness (i.e., the ability to obtain a large number of observations from a small number of subjects). For example, in some areas of physiological psychology, it might be common to obtain several hundred observations from each subject, and it is not unusual to see studies where the number of subjects ranges from 10 to 20, and sometimes samples are even smaller. These are not small sample studies in the traditional sense, because the number of observations is potentially huge (e.g., in some areas of vision research, studies in which 10 subjects each provide 400 trials are common). However, the small number of subjects does raise important concerns about the generalizability of the results. Even if true random sampling procedures are used, you have to be concerned about the possibility that the results obtained from one group of 10 subjects might be quite different from those obtained from another group of 10. In between-subject studies, they might not have the same set of worries. A random sample of 400 subjects is almost certain to provide results that generalize to the parent population, and a second random sample of 400 is almost certain to provide converging findings. In some cases where it is feasible to obtain hundreds of observations from each of a few subjects, it might be concluded that it is better to obtain a handful of observations from a larger set of subjects.

THE MULTIVARIATE ANALYSIS OF VARIANCE

Throughout, this book has concentrated on a family of statistical methods that all have one feature in common—that is, there is some specific variable, the dependent variable, that represents the focus of the analysis. In many studies, researchers are likely to collect data on several dependent variables, and are likely to conduct multivariate rather than univariate analyses. For example, suppose they are interested in determining whether there are differences in the outcomes of four specific training programs that have been proposed for teaching pilots to use a new global positioning technology. They might evaluate their performance in terms of the number of errors made in applying this technology, in terms of the amount of time needed to learn the technology, or in terms of increased efficiency in planning and sticking to flight routes. Rather than conducting three separate analyses of variance (i.e., one for errors, one for time spent, and one for flight planning), they are likely to use multivariate methods that combine these three dependent variables into a single analysis. Specifically, they are likely to carry out a multivariate analysis of variance (MANOVA).

There is a substantial literature dealing with power analysis in MANOVA (e.g., Maxwell & Delaney, 1990; Stevens, 1980, 1988, 2002; Vonesh & Schork, 1986). In a very general sense, the issues in determining the power of MANOVA are quite similar to those that affect the power of univariate ANOVAs. First, the power of MANOVA depends on *N*, the effect size, and the alpha level, just as in univariate ANOVA. Large samples, strong effects, and lenient alpha levels lead to high levels of power, whereas small samples, small effects, and stringent alpha levels lead to lower levels of power. Second, a principle already noted for repeated measures ANOVA also applies to MANOVA. In general, power is higher when the variables being examined (here, the multiple dependent variables that are combined to carry out the MANOVA) are highly correlated than when they are uncorrelated. Finally, the power of MANOVA depends on the number of dependent variables. A study that uses four or five dependent measures will tend to have more power than a study that uses one (leading to a univariate ANOVA) or two. Stevens (1988, 2002) presented useful tables for estimating power in MANOVA.

CONCLUSIONS

Multifactor ANOVAs involve asking several research questions at once, and the level of power often varies from question to question. Thus, when there are multiple independent variables, it is not enough to ask how much power there is in the analysis. Rather, researchers must ask how much power they have for each main effects test, how much power for each interaction, and so on. Designs that include repeated measures usually provide more power than equivalent between-subjects designs, because they allow them to control for sources of error that are difficult to eliminate in a between-subjects design.

The power of multivariate analyses of variance is influenced by many of the same factors that affect univariate ANOVAs. Power in MANOVA increases as the sample size increases, as the effect sizes get larger, and as alpha becomes less stringent. The power of MANOVA also depends on the number of dependent variables and the correlations between then. Power is higher when there are more dependent variables and when they are more highly correlated.

5 Illustrative Examples

The purpose of this chapter is to illustrate the application of power analysis. Using real examples from the published literature, it shows how the methods described in this book can be applied to simple statistical tests as well as to more complex experiments and field studies. In particular, it shows how the One-Stop *F* Table (Appendix B), together with the degrees of freedom tables presented in Appendices D and E, can be used to provide a range of useful information about the research and help in planning and interpreting research in the behavioral and social sciences.

The discussion begins with a detailed examination of two studies that used relatively simple statistical analyses (i.e., *t* tests and correlations), and shows how power analysis sheds light on the strengths and weaknesses, as well as the meaning of these studies. Next, it considers more complex statistical analyses, including factorial analyses of variance and multiple regression analysis, and shows how the tables developed in this book can be applied in analyzing these studies.

SIMPLE STATISTICAL TESTS

T Tests

Perhaps the most familiar statistical problem involves the comparison of the mean scores of subjects who receive some treatment with the mean scores of subjects who do not, via the *t* test. A research design in which subjects are randomly assigned to treatment and control conditions is very easy to implement and can provide very clear-cut and convincing answers to what are sometimes quite complex questions. To illustrate the many applications of statistical power analysis in studies employing this design, take a look at a study of the effects of regular exercise on placental growth rates in pregnant women (Clapp & Rizk, 1992).

Clapp and Rizk (1992) measured placental volumes in 18 healthy women who maintained a regular routine of exercise during pregnancy. Nine of the women engaged in aerobics, 4 ran, and 5 swam. The control group comprised 16 females who did not engage in such a regimen. Placental volumes were measured using modern ultrasound techniques at 16, 20, and 24 weeks gestation. Results of the study are reproduced in Table 5.1.

Two aspects of this study are especially noteworthy. First, the sample is small (i.e., $N = 34$). In most cases, this would mean very low power levels. However, in this study, the effects are quite strong. In the three time periods studied, exercise accounted for between 29% and 47% of the variance in placental volume (*d* ranges from 1.71 to 2.41). Because the apparent effects of exercise were quite substantial, it should be easy to rule out the hypothesis that the treatment has no effect (traditional null), or even that the true effects of exercise are at best small (minimum-effect hypothesis).

Traditional Versus Minimum-Effect Tests. This *t* test in this study has 32 degrees of freedom; if it is squared, the statistic is distributed as *F* with 1 and 32 degrees of freedom. The One-Stop *F* Table has entries for 1 and

TABLE 5.1

Placental Volumes Reported by Clapp and Rizk (1992)

Week	*Control (N = 16)*	*Treatment (N = 18)*	*t*	*F*	*PV*	*d*
16	106 (18)	141 (34)	3.68	13.55	.29	1.94
20	186 (46)	265 (67)	3.96	15.66	.33	1.71
24	270 (58)	410 (87)	5.45	29.66	.47	2.41

Note: Volumes are expressed in cm^3. Standard deviations are shown n parentheses. Note that *d* represents the mean difference divided by the control group *SD*. Use of pooled *SD* values yields somewhat smaller *d* values, but they will nevertheless exceed conventional benchmarks for "large" effects.

30 and for 1 and 40 degrees of freedom, but none for 1 and 32 degrees of freedom, meaning that it is necessary to interpolate to analyze these results. As can be seen in the One-Stop F Table, critical F values for the traditional null hypothesis at 1 and 30 degrees of freedom for $\alpha = .05$ and .01 are 4.16 and 7.56, respectively. The corresponding values at 1 and 40 degrees of freedom are as 4.08 and 7.31. Using Equation 9 in chapter 2, the $\alpha = .05$ and .01 critical values for 1 and 32 degrees of freedom when testing the traditional null hypothesis are 4.14 and 7.51. All of the observed F values in Table 5.1 are greater than 7.51, so the traditional null can be rejected at the .01 level and it may be concluded that regular exercise has *some* impact on placental volume.

The study conclusively rules out the possibility that exercise has no effect, and the data suggests that the actual effect is quite large. However, it is always possible that the true effect is small, and the large PV and d values observed here represent chance fluctuations in a process that usually produces only a small effect. You can use the One-Stop F Table to test the hypothesis that the effects of treatments are negligible, or perhaps small to moderate.

Again, using Equation 9 from chapter 2 to interpolate, the critical F ($\alpha = .05$) for the minimum-effect null hypothesis that the effect is small to medium in size (i.e., the null is that treatments account for no more than 5% of the variance in outcomes) is 9.52. All F values in Table 4.1 exceed 9.52, so the minimum-effect null hypothesis that the effects of treatments are no greater than small to moderate can be rejected with a 95% level of confidence. This hypothesis can also be rejected at the .01 level for mean placental volumes at 20 and 24 weeks (critical $F = 15.26$; observed $F = 15.66$ and 29.44 at 20 and 24 weeks, respectively). That is, there is 99% confidence that the effects of treatments exceed the definition of "small to moderate."

Power Estimation. Assume, as the results of this study suggest, that the effects of maternal exercise on placental volume are substantial (e.g., exercise accounts for 25% of the variance in placental volume). If this assumption is true, then this study possesses very high levels of power for rejecting the traditional null. With 1 and 32 degrees of freedom, an effect size of $PV = .25$ has an F equivalent value of 10.66. The critical F for achieving a power of .80 at $\alpha = .05$ in this study is 8.3. The power to reject the hypothesis that treatments account for less than 1% of the variance in the population is also well in excess of .80; the critical F for power of .80 in testing this minimum-effect hypothesis is 10.12.

In addition to placental volume, Clapp and Rizk (1992) examined a range of other dependent variables for the women in this study. If exercise could function as a viable treatment of fetal under- or over-growth, it might be expected to have some effect on the final birthweight of the babies as well. In this case, Clapp and Rizk (1992) reported the mean birth weight of babies for women in the control group as 3,553 gm (*SD* = 309) compared to 3,408 gm (*SD* = 445) for women in the treatment group. Reference to the One-Stop *F* Table shows that when testing hypotheses about birthweight, the study cannot even reject the traditional null hypothesis [F (1, 32) = 1.19, $p < .05$].

One possible explanation for this finding is low power. If the true effect of exercise on birthweight (as opposed to placental volume) is small, then power will be well below .50 in this study. For example, if it is assumed that the true effect of exercise on birthweight meets the conventional definition for a small effect (i.e., $PV = .01$), the *F* equivalent in this study is less than .32, whereas the critical *F* for a power of .50 to reject the traditional null at the .05 level is 4.03. Thus the power of this study to detect traditionally significant differences in birthweight is much less than .50 (about .09 in fact). Of course, the power to detect substantively meaningful differences is even lower.

Sample Size Estimation. How many subjects would be required to reliably detect differences in birthweight as a result of exercise during pregnancy? In order to answer this question, refer to Appendices D and E, which help in determining sample sizes needed, given that α = .05 and the desired level of power is .80. If it is assumed that the effect of exercise on birthweight is small ($PV = .01$, $d = .20$), then about 777 subjects in total (i.e., $N = df_{\text{hyp}} + df_{\text{err}}$; for the *t* test, $df_{\text{hyp}} = 1$), or about 389 subjects in each group (i.e., 777/2), will be needed to achieve this level of power in tests of the traditional null hypothesis.

The power analyses conducted here suggest that the Clapp and Rizk (1992) study was well suited for answering questions about the effect of exercise on placental volume, which was its major focus. Power exceeded .80 for tests of both traditional and minimum-effect null hypotheses. However, the sample size is quite inadequate for answering questions about the effects of exercise on birthweight. Because a small effect might reasonably be expected here, huge samples are probably needed to provide adequate power.

Correlations

McDaniel (1988) used very large samples to study the validity measures of preemployment drug use as predictors of job suitability in the military. Validity coefficients for preemployment use of drugs (e.g., marijuana, cocaine, various stimulants, and depressants) were calculated in samples ranging in size from 9,224 to 9,355 subjects. As might be expected, the correlations were all "statistically significant," and an incautious reading of these "significance" tests might mislead readers when arriving at conclusions about the value of these tests.

Several of the validities reported by McDaniel (1988) are shown in Table 5.2. These results are almost beyond the scope of the One-Stop *F* Table, but interpolating between $df_{\text{err}} = 1{,}000$ and $df_{\text{err}} = 10{,}000$ allows for the rejection of the traditional null hypothesis in every case at the .01 level. The interpolated critical *F* value for $\alpha = .01$ at 1 and 9,224 degrees of freedom is 6.64, which is much smaller than any of these observed *F* values. However, as noted later, the tests are significant because of the enormous samples involved, not because of the size of the effect.

Although significant in the traditional sense, McDaniel's (1988) validities are probably not meaningful in any substantive sense, and, as the author himself noted, employers would be better advised to base their employment decision on information other than applicant's previous history of drug use. This study is an excellent example of the pitfalls of relying on tests of the traditional null hypothesis. These correlations are all "significant,", and it is all too easy to confuse "significant" with "meaningful." As is noted later, these correlations are so small that they should probably be ignored altogether.

Traditional Versus Minimum-Effect Tests. An examination of Table 5.2 shows just how small these validities are, accounting for less than one half of a percent of the variance in job suitability in each case. This fact is reflected in the One-Stop *F* Table as well. At 1 and 9,224 degrees of freedom, the interpolated critical *F* for rejecting the null hypothesis that the effect is negligibly small (i.e., 1% or less of the variance) is 126.1, well above any of the observed *F* values. In other words, none of the validities reaches the level designated as "trivial." Although

TABLE 5.2

Predictive Validity of Preemployment Drug Use Reported by McDaniel (1988)

Drug	*N*	*Validity*	*PV*	*t*	*F*
Marijuana	9,355	.07	.0049	6.79	46.06
Cocaine	9,224	.04	.0016	3.85	14.81
Stimulants	9,286	.07	.0049	6.76	45.73
Depressants	9,267	.07	.0049	6.76	45.63

they are significantly different from zero, it is clear that the relation between these tests and job suitability is so weak that the variables could be treated as virtually independent.

This example illustrates an important feature of minimum-effect null hypotheses. Unlike the traditional null, a minimum-effect null cannot necessarily be rejected simply by collecting more data. If an effect is genuinely negligible, it will continue to be negligible no matter how many subjects are tested. This phenomenon is reflected in the One-Stop Table by the fact that critical *F* values for the 1% and 5% minimum-effect nulls do not asymptote as df_{err} increases. As you can see, as df_{err} gets very large, the critical values for minimum-effect nulls keep pace with the increase and do not allow for rejection of the minimum-effect null hypothesis unless the effect is genuinely meaningful. Herein lies another important advantage of minimum-effect null hypothesis testing over traditional null hypothesis testing. When testing a minimum-effect null hypothesis, rejection cannot be guaranteed simply by collecting a large enough sample.

Power Estimation. As noted earlier, the large samples employed in this study give tremendous power in tests of the traditional null hypothesis. If it is assumed that the relation between preemployment tests and job suitability is only one half as large as the conventional definition of a small effect (i.e., $PV = .005$), then power would still be in excess of .80 for all of the tests reported in Table 4.2. With $N = 9{,}224$ and an assumed effect of $PV = .005$, the *F* equivalent is 46.35, whereas the critical *F* for achieving power of .80 is 7.81. It is little wonder that the author of this study rejected the traditional null hypothesis. Even if the true effect is absolutely trivial, there is still plenty of power for rejecting the hypothesis that there is no effect whatsoever.

In contrast, if the hypothesis to be tested is that these tests account for 1% or less of the variance in suitability, McDaniel's study does not achieve power of even .50; to reach this level of power for samples this large would require an *F* of over 120. As noted earlier, if the samples included 90,000 rather than 9,000 individuals, researchers still would not have power for rejecting a minimum-effect null with such data. The effect is much too small to be sensibly described as "meaningful," and no matter how large the sample, it should still be small. The fact that a study can never attain the power needed to reject a minimum-effect hypothesis when the effect is in fact negligible is neatly illustrated with this study.

STATISTICAL TESTS IN COMPLEX EXPERIMENTS

ANOVA

In a meticulously designed study, Martin, Mackenzie, Lovegrove, and McNicol (1993) examined the effect of tinted lenses on the reading performance of children with specific reading disabilities (SRDs). This treatment was first suggested by Irlen (1983), who distinguished a condition called "scotopic sensitivity syndrome" that could be ameliorated to some degree by prescription of lenses of certain tints. Controversy surrounds this treatment because it seems highly unlikely to many psychologists that filtering a few wavelengths from the perceptual array could have any impact on the higher cognitive processes associated with reading (Evans & Drasdo, 1991). Irlen (1983) proposed a mechanism involving perceptual distortions at certain wavelengths on the retina, but other authors have attributed any improvement in reading to placebo effects (Cotton & Evans, 1990; Winter, 1987).

From an original sample of 300, Martin et al. (1993) selected 60 subjects for further study. Twenty subjects were normal readers, 20 were SRDs who could be treated by tinted lenses, and 20 were SRDs deemed unsuitable for the treatment. The sample selection process controlled for such variables as native language, eyesight, intelligence, behavioral disorders, organic problems, and exposure to previous therapies. The three groups were compared on three occasions (pretest, posttest, and follow-up) using the Neale Analysis of Reading Ability (Neale, 1973), which provided measures of reading accuracy and comprehension. A posttest occurred one year after the pretest and follow-up occurred 6 months later. Although the sample in this study was small, the authors were able to achieve substantial sensitivity because of their careful procedures and because they used a repeated measures design, in which multiple observations were obtained from each subject.

The data in this study were analyzed using a 3 x 3 analysis of variance. Martin et al. (1993) reported a significant main effect for Groups on accuracy [F (2, 57) = 65.57] and comprehension [F (2, 57) = 7.72]. The main effect for Occasion was also significant in each case [F (2, 114) = 24.99 for Accuracy and F (2, 114) = 36.43 for Comprehension]. The One-Stop F Table contains a range of useful information that helps put these results in perspective.

Power Estimation: Main Effects. There are good reasons to believe that the main effects in this study are large, and if this is true, the power of all main effect tests was extremely high. For the Groups main effect, the critical *F* value for achieving power of .80 to reject the traditional null (α = .05) is 5.06. The *F* equivalent for a large effect in this study is 9.50, suggesting power well in excess of .80. Power was also greater than .80 for rejecting (at the .05 level) the minimum-effect null hypothesis that main effects account for 1% or less of the variance in treatments. Here, the critical *F* value was 6.19. However, in tests of the Groups main effect, power is lower than .80 for tests of the minimum-effect null for a small to medium effect (i.e., that treatments account for less than 5% of the variance in the population; the critical *F* = 9.76).

The same conclusions can be drawn about the Occassions main effect. In this case, the critical *F* value for achieving power of .80 to reject the 1% minimum-effect null hypothesis at the.05 level was 6.96; for tests of the 5% minimum-effect null, the critical value is 12.72. With $df_{\text{hyp}} = 2$ and $df_{\text{err}} = 114$, a large effect (i.e., *PV* = .25) yields an *F* equivalent value of 19.00, which is well in excess of the critical *F* needed to attain power of .80 for tests of the minimum-effect null hypotheses described earlier.

Testing Interaction Hypotheses. Examination of the main effects does not really address the principal research question (i.e., the effects of tinted lenses on reading performance of children with SRDs). If tinted lenses work, then the study should observe performance of the treated SRDs to be more similar to the untreated SRDs in the pretest condition but more similar to the subjects without reading disorders during posttest, hopefully persisting until follow-up. This is an interaction hypothesis. The interaction of groups by occasion was significant in the traditional sense for Accuracy [*F* (4, 114) = 2.98] and Comprehension [*F* (4, 114) = 3.13]. In both cases, the critical *F* (α = .05) value was 2.45 for tests of the traditional null. A minimum-effect null (i.e., that the interaction effect accounts for 1% or less of the variance: critical *F* = 3.10) can also be rejected for comprehension measures.

The power of tests of these interactions can be determined from the One-Stop *F* Table. Suppose, for example, that the interaction effect is moderately large. If the population *PV* = .10, the *F* equivalent for $df_{\text{hyp}} = 4$, $df_{\text{err}} = 114$ is 3.16. The corresponding critical value for a power of .80 is 3.07. Power to reject the 1% minimum-effect null is somewhat lower. In this case, critical *F* values are 2.27 and 3.90 for power levels of .5 and .8, respectively. The *F* equivalent for a moderate effect (i.e., *PV* = .10) falls between these critical values, so Equation 8 (chap. 2) can be used to obtain a more accurate estimate. The power to reject the 1% minimum-effect null is approximately .5 + .3 (3.16 – 2.27) / (3.90 – 2.27) = .66. Thus, although acceptable for testing the traditional null hypothesis, this study is not quite adequate for addressing minimum-effect null hypotheses.

There is a reasonable level of power for detecting moderately strong interactions. Suppose, however, that the interaction effect is small (e.g., $PV = .01$). With $df_{\text{hyp}} = 4$ and $df_{\text{err}} = 114$, the equivalent F value is now .287, which is obviously too small to yield even marginal levels of power. In other words, the sample in this study is large enough to give adequate power if and only if the true interaction is at least moderately strong.

In a complex experiment, each statistical test may show a different level of power. First, the predicted and/or actual effect sizes may differ for main effects, simple interactions, and higher order interactions. Second, the degrees of freedom for different tests in a model may vary, sometimes substantially. In general, the study is most likely to have higher power for testing main effects and lower power for testing complex interactions. First, main effects usually have more degrees of freedom. Second, main effects are usually (but not always) stronger than complex interactions. When the focus of the study is a complex three- or four-way interaction, very large samples might be needed to provide adequate power.

Martin et al. (1993) concluded that tinted lenses have no effect on reading ability, but the significant interactions just discussed do not warrant such a conclusion. The comprehension interaction, in particular, shows that treated SRDs performed like untreated SRDs in the pretest but improved to almost the same level as normals in the posttest and follow-up, after lenses were applied.

Potential Difficulties in Estimating Power in Repeated Measures Designs. This study involved obtaining multiple measures from each of several subjects, and both hypothesis testing and power analysis in repeated measures designs can be complicated, because these designs lead to violations of important statistical assumptions that underlie the analysis of variance. In particular, research designs that involve obtaining multiple measures from each respondent can lead to violations of assumptions of independence of sphericity (i.e., the assumption that the variances of the differences between all possible pairs of repeated measures are equal). To obtain accurate results in such designs, both the degrees of freedom and an estimate of the noncentrality parameter must be adjusted by a factor "epsilon" (ε), which reflects the severity of violations of the assumption of sphericity [e.g., the best estimate of λ in repeated measures designs is given by ($\varepsilon * df_{\text{err}} * PV/(1 - PV)$); see Algina & Keselman, 1997, for discussions of sphericity and power]. However, this correction factor is not always reported, and cannot usually be calculated on the basis of results likely to be reported in a journal article.

A conservative approach to this problem is to make a worst-case correction (Greenhouse & Geisser, 1959). The factor ε ranges in value from a maximum of 1.0 (indicating no violation) to a minimum of $1/(k-1)$, where k represents the number of levels of the repeated measures factor. When the degrees of freedom for factors involving repeated measures are multiplied by epsilon, it is possible to obtain a conservative test of significance by comparing the obtained F to the critical value of F using the epsilon-adjusted degrees of freedom.

For example, in this study, data are collected from each subject on three occasions, meaning that Occasion is a repeated measures factor, and the Group X Occasion factor has a repeated measures component. The worst-case estimate of epsilon is that $\varepsilon = .5$ [i.e., epsilon = $1/(3-1)$]. To use this worst-case estimate ε here, multiply the degrees of freedom for the Occasion effect and the Occasion X Group effects by .5 (i.e., use degrees of freedom of $df_{\text{hyp}} = 1$, $df_{\text{err}} = 57$ and $df_{\text{hyp}} = 2$, $df_{\text{err}} = 114$, respectively, to test the Occasion and Occasion X Group effects rather than the actual degrees of freedom of $df_{\text{hyp}} = 2$, $df_{\text{err}} = 57$ and $df_{\text{hyp}} = 4$, $df_{\text{err}} = 114$). Similarly, adjust the estimate of λ, also multiplying the estimate of λ by .5.

In practice, a good estimate of power can be obtained by simply multiplying both the degrees of freedom and *PV* by this worst-case estimate of epsilon (rather than directly adjusting your estimate of λ). A comparison of the power estimate obtained without any epsilon correction (the power estimates in this example did not include any correction for violations of the sphericity assumption) with the power estimate obtained making a worst-case assumption about violations will provide a pretty good idea of the range of power values that would be reasonable to expect.

Multiple Regression

Bunce and West (1995) examined the role of personality factors (propensity to innovate, rule independence, intrinsic job motivation) and group climate factors (support for innovation, shared vision, task orientation, participation in decision making) in explaining innovation among health service workers. They carried out a 17-month, three-stage longitudinal study, and like most longitudinal studies, suffered significant loss of data (subject mortality) across time. Their N dropped from 435 at Stage 1 to 281 and 148 at Stages 2 and 3. Incomplete data reduced the effective sample size for several analyses further; several critical analyses appear to be based on a set of 76 respondents.

One analysis used Stage 1 innovation, personality factors, and group factors as predictors in a hierarchical regression model, where the dependent variable was innovation at Stage 2. The results of this analysis are summarized in Table 5.3.

TABLE 5.3

Hierarchial Regression Results Reported in Bunce and West (1995)

Predictor	R^2	*df*	*F*	ΔR^2	*df*	*F*
Innovation—Stage 1	.18	1,75	16.73*			
Personality	.33	4,72	8.86*	.15	3,72	5.37*
Group Climate	.39	8,68	5.38*	.06	4,68	1.67

* $p < 05$ in tests of the traditional null hypothesis

The principal hypotheses tested in this analysis were that personality accounts for variance in Stage 2 innovation not explained by innovation at Stage 1, and group climate accounts for additional variance not accounted for by personality and Stage 1 innovation. Because Stage 1 innovation is entered first in the equation, tests of the changes in R^2 as personality and climate are added to the equation must be interpreted as reflecting the influence of these factors on changes in innovation over time. The results suggest that personality has a significant effect, but that group climate does not account for variance above and beyond Stage 1 innovation and personality.

Power Estimation The results in Table 5.3 show a pattern encountered in some previous examples, a relatively small sample combined with some reasonably large effects (e.g., Stage 1 innovation accounts for 18% of the variance in Stage 2 innovation), which makes it difficult to determine offhand whether or not the study will have enough power for its stated purpose. Some quick calculations suggest that it does not posses sufficient power to test all of the hypotheses of interest.

Rather than starting from the reported R^2 and F values, suppose the study used standard conventions for describing large, medium, and small effects to structure the power analysis. In hierarchical regression studies, the predictors are usually chosen to be relevant to the dependent variable (which means they should each be related to Y), and are therefore usually also intercorrelated (i.e., several variables that are all related to Y are likely to also be related to one another). As a result, the study will generally find that the first variable entered will yield a relatively large R^2, and R^2 will not increase as quickly as more variables are entered (J. Cohen & P. Cohen, 1983). This might lead researchers to expect a large effect for the first variable entered, a smaller change in R^2 for the next variable, and a small change in R^2 for the last variable. Chapter 2 noted that R^2 values of .25, .10, and .01 corresponded to conventional definitions of large, medium, and small effects (See Table 2.2). These values turn out to be reasonably similar to the actual R^2 and ΔR^2 values shown in Table 4.3 (i.e., .18, .15, and .06, respectively). Even more to the point, the overall R^2, which represents sum of these conventional values (i.e., $R^2 = .36 = .25 + .10 + .01$), is very similar to the actual overall value (i.e., $R^2 = .39$) reported by Bunce and West (1995).

To estimate power for detecting R^2 values of .25, .10, and .01, given the degrees of freedom in this study, first translate the R^2 values into F equivalents, using equations shown in Table 2.1. The F equivalents are 25.0, 3.69, and .27, respectively. There is plenty of power for testing the hypothesis that $R^2 = .25$ [$F(1, 75) = 25.0$]; the critical tabled F for this level of power is 8.01. Even when testing the minimum-effect hypothesis that the first variable entered accounts for 5% or less of the variance, power far exceeds .80.

There is also a reasonable level of power for testing the hypothesis that the second variable entered into the regression equation will have a medium effect [i.e., $R^2 = .10$, $F(3,72) = 3.69$]. Interpolating between values in the One-Stop F Table, the study would achieve power of .80 with an F value of 3.79. The F here is quite close, and the power of this test is approximately .78.

If it is assumed that the third in a set of intercorrelated predictors will generally yield a small increment in R^2 (i.e., $\Delta R^2 = .01$), there is clearly not enough power to test that hypothesis. To achieve power of .50, the study would need an F of 1.68; assuming a small effect here, F is only .27.

The conclusions reached by looking at these three conventional values closely mirror those that would be obtained if the actual R^2 values were used. The F values for the first predictor (Stage 1 innovation), second predictor (personality), and third predictor (group climate) are 16.73, 5.37, and 1.67, respectively. Power easily exceeds .80 in tests of the hypothesis that stage 1 innovation is related to Stage 2 innovation (the critical F for this level of power is 8.01). Power also exceeds .80 for testing the hypothesis that personality accounts for variance not explained by Stage 1 innovation (the critical F is 3.79; the observed F is 5.37). Power is just below .50 for testing the hypothesis that group climate accounts for variance not accounted for by the other two predictors (the critical F is 1.68; the observed F is 1.67).

Sample Size Estimation. The sample is certainly large enough to provide a powerful test of the hypothesis that Stage 1 innovation predicts Stage 2 innovation. Power is also reasonably high for testing the hypothesis that adding personality to the equation yields a significant increase in R^2. However, a *much* larger sample would be needed to provide a powerful test of the last hypothesis (i.e., that group climate explains additional variance).

A look down the $df_{\text{hyp}} = 4$ column of the One-Stop F Table reveals that even when $df_{\text{err}} = 10{,}000$, the critical F value needed to have power of .80 is larger than 1.67. If researchers are seriously interested in pursuing hypotheses of this sort, truly enormous samples are necessary.

CONCLUSIONS

This chapter has provided numerous illustrations of how power analysis can be used to help in designing and interpreting research studies. In particular, use of the One-Stop F Table in interpreting existing studies was demonstrated. The use of Appendices D and E in estimating sample sizes required to achieve desired levels of power was also illustrated.

This chapter began by looking at a small sample study of the effects of exercise during pregnancy and noted that the study was adequate for answering questions about some dependent variables (due to the large effect of exercise), but inadequate for answering questions about dependent variables where a small or even moderate effect might be expected. Next, it looked at a large sample study of the effect of preemployment drug use on job suitability and noted that, although statistically significant, the effects were trivially small. The futility of trying to interpret such results was also noted.

A tightly controlled study of the effect of tinted lenses on the reading performance of children with specific reading disabilities illustrated the applications of statistical power analysis in a complex ANOVA. The discussion lamented the fact that the authors did not report power analysis information, especially given that they really wanted to accept the null hypothesis (i.e., that using the lenses has no real effect). It identified trends in their data that were inconsistent with their conclusions, and suggested that the use of procedures designed to control for Type I errors may have caused more problems than they could possibly have solved.

Finally, the use of statistical power analysis was illustrated in a study that employed hierarchical regression techniques. This study examined the role of individual and group factors in influencing innovation; the authors suggested that group factors made no independent contribution to the prediction of innovation. The analysis suggests that the study had woefully inadequate power for that purpose, and further that very large samples would be needed to provide adequate power. As a result, the authors' failure to reject the traditional null hypothesis cannot be taken as evidence one way or another about the incremental effects of group climate on innovation.

6 The Implications of Power Analyses

The power of a statistical test is the probability that the study reject the null hypothesis being tested, given that this null hypothesis is in fact wrong. As noted throughout, the traditional null hypothesis that treatments have no effect whatsoever (or that the correlation between two variables is precisely zero, or any other hypothesis of "no effect") is very often wrong, and in this context the statistical power of a test is essentially the probability that the test will lead to the correct conclusion. When testing the traditional null hypothesis, it is obvious that power should always be as high as possible. When testing a minimum-effect hypothesis (e.g., that the effect of treatments is negligibly small, but not necessarily precisely zero), the implications of varying levels of statistical power are potentially more complex, and a wider range of issues needs to be considered in determining how to use and interpret statistical power analysis.

This chapter begins with a discussion of the implications of statistical power analysis for tests of both traditional and minimum-effect null hypotheses. Next, it discusses the benefits of taking statistical power seriously. Some of these are direct and obvious (e.g., if researchers do a power analysis, they are less likely to conduct a study in which the odds of failure substantially outweigh the odds of success), but there are also a number of indirect benefits to doing power analyses that may, in the long run, be even more important than the direct benefits. Finally, it considers the question of whether power analysis renders the whole exercise of testing the traditional null hypothesis moot. If power is very high (or very low), then the outcome of most statistical tests is not really in doubt, and the information from these tests might be severely limited as power reaches either extreme.

TESTS OF THE TRADITIONAL NULL HYPOTHESIS

Chapter 1 noted that two types of errors might be possible in testing a statistical hypothesis. First, researchers might reject the null hypothesis when it is in fact true (a Type I error). Second, they might fail to reject the null hypothesis when it is in fact wrong (a Type II error). Textbooks invariably stress the need to balance one type of error against the other (e.g., procedures that minimize Type I errors also lead to low levels of power), but when the null hypothesis is almost certain to be wrong there is little to be gained and much to be lost by attempting to maintain such a "balance" (Murphy, 1990).

The fact that the traditional null hypothesis is so often wrong leads to three conclusions about statistical power: (a) A study cannot have too much power, (b) studies should take the simplest and most painless route to maximizing power, and (c) tests with insufficient power should never be done.

You Cannot Have Too Much Power

If the null hypothesis is wrong, researchers cannot make a Type I error, and the only possible way they can make an error in statistical hypothesis testing is to fail to reject H_0. The reason for repeating this point so many times is that it flies in the face of convention, where substantial attention is often devoted to the unlikely possibility that a Type I error might occur. In tests of the traditional null, power is essentially the probability that the test will reach the right conclusion (because the traditional null is usually wrong), and there is no statistical rationale for arguing that power should not be high. There are many practical problems with attaining high levels of power, as noted later. However, it is always advantageous to maximize power in tests of the traditional null hypothesis.

Maximizing Power: The Hard Way and the Easy Way

There are two practical and eminently sensible ways to attain high levels of power. The easy way is to change the alpha level. As shown in chapters 2 and 3, power is higher when a relatively lenient alpha level is used. Traditionally, the choice of criteria for defining statistical significance has been between alpha levels of .05 and .01. When testing the traditional null, there is little scientific or statistical advantage to using a stringent test, and the alpha level for tests should generally be set as high as possible. Unfortunately, you are likely to meet resistance if they use alpha levels of .10, .20, or anything other than .05 or some other conventional figure, but this resistance is misplaced. Higher alpha rates yield more power, often with no meaningful increase in the likelihood of a Type I error.

The second strategy for maximizing power is to increase the sensitivity of the study, which generally implies using larger samples. Even though this strategy is more demanding than simply changing the alpha level, it is strongly recommended. Large, carefully constructed samples increase the generalizability and stability of findings, and they decrease the possibility that sampling error will lead to meaningless or misleading results. There is more about this in later sections.

Whereas it is clearly harder to increase power by increasing N than by increasing α, this strategy has the immense benefit of improving a study. Simply changing the alpha level does nothing to enhance the meaningfulness or interpretability of the research, but the use of large samples helps to minimize one of the recurring problems in social science research—the overreliance on the unstable results obtained in small samples (Schmidt, 1992).

Tests with Insufficient Power Should Never Be Done

Suppose someone is diagnosed with an ulcer, and the doctor told the person about a new treatment. This treatment is more likely to make things worse than to make things better, and there are alternative treatments available that do not have this problem. Would you try the new treatment ? Our answer is "no," and we believe this analogy applies exactly to statistical tests of the traditional null hypothesis. If power is low, do not carry out a test of the traditional null hypothesis.

When power is less than .50 and it is virtually certain that H_0 is wrong, the test is more likely to yield a wrong answer than a right one. More to the point, the test is unlikely to produce new and useful knowledge; it is more likely to mislead. If researchers were certain before the test that H_0 is false, a test that rejects H_0 does not tell anything that is not already known. A test that fails to reject H_0 should not change anyone's mind either (if H_0 is wrong virtually by definition, the results of the test should not change this), but people will sometimes be mislead by their data. Low power tests are unlikely to have any effect except to mislead and confuse researchers and readers.

TESTS OF MINIMUM-EFFECT HYPOTHESES

The alternative to testing the traditional null hypothesis that treatments have no effect is to test the minimum-effect null hypothesis that the effect of treatments is so small that it could be safely ignored. Different disciplines or research areas might require substantially different operational definitions of a "negligibly small" effect, and the standards suggested in previous chapters and in the One-Stop *F* Table (i.e., treatments accounting for less that 1%, or in some cases less than 5%, of the variance in outcomes have negligibly small effects) will not always apply. Nevertheless, tests of minimum-effect null hypotheses are necessary if the whole enterprise of statistical hypothesis testing is to prove useful.

Statistical power analysis can be used to its fullest advantage in tests of minimum-effect null hypotheses. Because the hypothesis being tested may very well be true, it becomes important to develop specific procedures and criteria for "accepting the null," or determining when the evidence is consistent or inconsistent with the proposition that the effects of a particular treatment are indeed negligible; power analysis is extremely useful for this purpose. It also becomes important to give serious consideration to an issue that is usually (and incorrectly) presented in the context of traditional null hypothesis tests (i.e., the appropriate balance between Type I and Type II errors).

Accepting the Null

In traditional null hypothesis testing, the idea of accepting the null hypothesis is sometimes treated as a sort of heresy. Rather than allowing researchers to accept the hypothesis that treatments have no effect, the traditional framework usually leaves them with two options: deciding that there is sufficient evidence to reject the null (i.e., a significant outcome), and deciding that there is not yet enough evidence to reject the null (i.e., a nonsignificant result). Because researchers already know that the traditional null is almost certain to be false, the fact that they have not yet accumulated enough evidence to confirm this fact tells more about the study than about the substantive phenomenon being studied.

As noted throughout, power is substantially affected by the size of the sample. If *N* is very small, researchers will not reject the null, no matter what research question they are pursuing. If *N* is large enough, they will reject the traditional null, again no matter what research question they are pursuing. It is hard to resist the conclusion that tests of the traditional null hypothesis are little more that indirect measures of the sample size! In tests of the traditional null, the most logical interpretation of a nonsignificant result is that the sample is too small.

Occasionally, the door is left open for treating nonsignificant results as meaningful. For example, some journals allow for the possibility of publishing nonsignificant results, at least under some conditions (e.g., the inaugural issue of *Human Performance* included an editorial suggesting that nonsignificant results would be treated as meaningful if specific research design criteria, including a demonstration of adequate statistical power, were met). The argument that is sometimes offered is that if well-designed studies fail to detect an effect, then this might provide some evidence that the effect is likely to be a very small one, and the null hypothesis might be very close to being true, even if the effect of treatments is not precisely zero. Recently, Bayesian approaches have been applied to the problem of statistically demonstrating that an effect is so small that it should be effectively ignored (Rouanet, 1996). Nevertheless, the bias against "accepting the null" runs so strong in tests of the traditional null hypothesis that this framework simply does not leave any appealing alternative when the effect of treatments is negligibly small. Researchers will either collect a very large sample and reject the null (which may mislead them into thinking that the effect of treatments is something other than trivial) or they will fail to reject it and perhaps go out and collect more data about an essentially meaningless question.

In tests of minimum-effect hypotheses, there is a realistic possibility that the hypothesis being tested (e.g., that the effect of treatments is at best negligible) is indeed true, and there is a real need to develop procedures or conventions for deciding when to "accept the null." Power analysis plays a critical role in determining and defining those procedures or conventions.

Suppose the hypothesis being tested is that the effect of treatments is at best negligible (e.g., treatments account for 1% or less of the variance in outcomes). A powerful study could provide strong evidence that this hypothesis is in fact true. For example, if power is .80, then this translates into odds of 4 to 1 that a statistical test will reject this hypothesis if it is in fact false. Failure to reject the null under these conditions can mean only one of two things: the null really is true, or the null is false, and this is that one test in five that yields the wrong result. The most logical conclusion to reach in this study is that the effects are negligibly small.

As noted in chapter 3, a complete evaluation of the meaning of the outcomes of statistical tests requires some knowledge about the probability that the null hypothesis being tested actually is true (i.e., the prior probability of H_0). The central weakness of traditional null hypothesis testing is that this prior probability is thought or defined to be vanishingly small, and perhaps zero (Murphy, 1990). If this prior probability is zero, then tests of the null hypothesis cannot provide much useful information.

The central weakness of the alternative approach described in this book, in which minimum-effect hypotheses are framed and tested, is that this prior probability is generally unknown. In chapter 3 noted that the prior probability of the traditional null hypothesis would necessarily be very low, and by most definitions is zero. It is likely that the prior probability of a minimum-effect null will also be somewhat low, especially if the study is testing the hypothesis that the effect of treatments is at best negligible. The rationale for believing that this prior will be small is that most treatments in the social and behavioral sciences do indeed have an effect. Chapter 1 noted Lipsey and Wilson's (1993) review of over 300 meta-analyses of research studying the efficacy of psychological, educational, and behavioral treatments. These meta-analyses summarize the results of thousands of studies in areas ranging from smoking cessation success rates to the effectiveness of computer-aided instruction. Over 85% of the meta-analyses they summarized reported effects that exceeded conventional criteria for "small effects" (i.e., d = .20 or PV = .01 or less). It is clear from the massive body of research summarized in Lipsey and Wilson that a broad range of treatments and applications in the social and behavioral sciences have at least some effect, and the likelihood that a new treatment will have an absolutely negligible effect seems small, especially if the new treatment is solidly based in theory and research.

Balancing Errors in Testing Minimum-Effect Hypotheses

In tests of the traditional null hypothesis, Type I errors are practically impossible and there is virtually nothing to be lost by setting alpha as high as possible. In tests of the minimum-effect hypothesis, the strategy of setting alpha as high as possible is no longer appropriate. The whole distinction between traditional and minimum-effect null hypotheses is that there is some realistic possibility that a minimum-effect null is true, and it is therefore possible to make a Type I error.

Although Type I errors are a real possibility in tests of minimum-effect null hypotheses, this does not mean that power should be ignored in carrying out these tests. The possibility of Type I errors should not blind researchers to the substantial likelihood that they will make Type II errors if they choose an unduly stringent alpha level. Choose any cell of the One-Stop F Table in Appendix B, and it becomes evident that (a) a larger F value is needed to reject the a minimum-effect null than to reject the traditional null, given the same alpha level, and (b) a larger F value is needed to reject the hypothesis that effects are small to moderate (e.g., they account for 5% or less of the variance) than to reject the hypothesis that these effects are negligibly small (i.e., they account for 1% or less of the variance). That is, all other things being equal, it is harder to reject a minimum-effect hypothesis than to reject the hypothesis that treatments have no effect whatsoever. The more demanding the hypothesis (e.g., 5% vs. 1% as the upper bound to be tested), the harder it is to reject H_0. So, there is usually no good reason to make things even more difficult than they already are by choosing an unrealistically stringent alpha level. Earlier, it was suggested that when testing the traditional null hypothesis, the .01 level should usually be avoided. The same advice holds for tests of minimum-effect hypothesis; the .01 alpha level should still be avoided in most cases.

A two-part test may be used to determine whether to use an alpha level of .01 rather than .05. First, .01 makes sense only if the consequences of a Type I error are relatively serious. Chapter 3 discussed concrete ways of comparing the perceived seriousness of Type I and Type II errors, and noted that researchers often act as if falsely rejecting the null is much more serious than failing to reject the null when necessary. This may well be true in some settings, but before deciding to set a stringent alpha level (thus markedly decreasing power), explicitly consider the relative costs of the two errors. Choose .01 only if there is a good reason to believe that a Type I error is substantially more serious than a Type II error.

Second, .01 makes sense only if the prior probability that the null hypothesis is true is reasonably high. That is, if the hypothesis to be tested is that a treatment has at most a negligible effect (e.g., it accounts for 1% or less of the variance in outcomes), the study should be concerned with Type I errors only if there is some realistic possibility that the effects of treatments are indeed trivial. Finally, keep in mind that this is a two-part test. Use the .01 level rather than the .05 level only if the consequences of a Type I error are large and the possibility that one might actually occur are substantial. In all other cases, use .05 as an alpha level, and use an even more lenient alpha level whenever possible.

POWER ANALYSIS: BENEFITS, COSTS, AND IMPLICATIONS FOR HYPOTHESIS TESTING

If power analysis is taken seriously, there will be fundamental changes in the design, execution, and interpretation of research in the social and behavioral sciences. Most of these changes should be beneficial, and this discussion advocates power analysis. As noted later, the indirect benefits of power analysis may prove, in the long run, even more important than the direct benefits of adopting this approach. There are, of course, some costs associated with incorporating power analysis in the design and interpretation of research; however, the benefits still substantially outweigh the costs. Finally, it is useful to consider the implications of having extreme levels of power (either extremely high or extremely low) when conducting statistical hypothesis tests.

Direct Benefits of Power Analysis

As noted in chapter 1, power analysis can be used as both a planning tool (e.g., determining how many subjects should be included in a study) and a diagnostic tool (e.g., making sense out of previous studies that have either reported or failed to report significant results). Individuals who incorporate statistical power analysis into their research repertoire are better equipped to both plan and diagnose research studies, and they directly benefit from the information provided by power analyses.

Planning Research Statistical power analysis provides a rational framework for making important decisions about the design and scope of a study. To be sure, there are many subjective decisions that must be made in applying power analysis (e.g., what effect size is anticipated, what alpha level is best), and the techniques described in this book do not represent a foolproof formula for making decisions about research design (e.g., choosing between repeated measures or between-subjects designs) or sample size. The advantage of power analysis over other methods of making these important decisions, which are often made on the basis of force of habit, or by following the lead of other researchers, is that it makes explicit the consequences of these design choices for a study. If researchers are seriously interested in rejecting the null hypothesis, they think a power analysis is absolutely necessary in making good choices about study design and sample size.

Power analysis also highlights the importance of a decision that is usually ignored, or made solely on the basis of conventions in one's field (i.e., the alpha level that defines statistical significance). The choice of stringent criteria (e.g., $\alpha = .01$) is sometimes interpreted as scientifically rigorous, whereas the choice of less rigorous criteria (e.g., $\alpha = .10$) is sometimes derided as "soft science." Nothing could be farther from the truth. In fact, any decision about alpha levels implies some wish to balance Type I and Type II errors, and power analysis is absolutely necessary in order to make any kind of sense out of that balance. Once researchers appreciate the implications of choosing different alpha levels for the statistical power of their studies, they are more likely to make sensible choices about this critical parameter.

If power analysis is taken seriously, it is likely that fewer studies with small samples or insufficient sensitivity will be done. Thus, researchers benefit substantially by knowing whether the study they have in mind has any real likelihood of detecting treatment effects. As noted later, the indirect benefits to the field as a whole that might come with a decline in small sample research are even greater than the benefits to the individual researcher.

Interpreting Research. One criticism of tests of the traditional null hypothesis is that they can mislead researchers and readers. Researchers who uncover a significant result are likely to confuse that with an important or meaningful result. This is hardly surprising; most dictionary definitions of "significant" include "importance," "weight," or "noteworthy" as synonyms. Similarly, "nonsignificant" is easily confused with "not important" or "nonmeaningful." As power analysis clearly shows, very meaningful and important treatment effects are likely to be nonsignificant if the study lacks power, whereas completely trivial effects are likely to be significant if enough data are collected. It is impossible to sensibly interpret significant or nonsignificant results without considering the level of statistical power in the study that produced those results.

To give a concrete illustration, suppose a researcher reviewed a dozen studies, all of which reported a nonsignificant correlation between attitudes toward drug use and subsequent drug consumption. What does this mean? If the studies are all based on small samples, it is entirely possible that there is a real and meaningful correlation between attitudes and subsequent behavior (e.g., if $N = 30$ and $\alpha = .05$, power for detecting a correlation as large as .30 is only .50), and the studies simply did not have enough power to detect it. On the other hand, if all of the studies included very large samples (e.g., $N = 2{,}500$), it could probably be concluded that there is essentially no relation between present attitudes and future behavior. Although the traditional null hypothesis might not be literally true in this instance, it would have to be be very nearly true. With this much power, the studies reviewed would have almost certainly detected any consistent relation between attitudes and behavior.

Indirect Benefits of Power Analysis

The widespread use of power analysis is likely to confer many indirect benefits. Most notably, studies designed with statistical power in mind are likely to use large samples and sensitive procedures. Perhaps even more important, power analysis directs the researcher's attention toward the most important parameter of all (i.e., the effect size). The ultimate benefit of statistical power analysis may be that it forces researchers to think about the strength of the effects they study, rather than thinking only about whether or not a particular effect is significant.

Large Samples, Sensitive Procedures. Small samples are the bane of social science research (J. E. Hunter & Schmidt, 1990; Schmidt, 1994). These studies produce unstable results, which in turn produce attempts to develop theories to "explain" what may be little more than sampling error. If power analyses were routinely included in the process of designing and planning studies, large samples would be the norm and sampling error would not loom so large as a barrier to cumulative progress in research.

Proponents of meta-analysis (e.g., Schmidt, 1994) note that by combining the outcomes of multiple small sample studies, it is possible to draw sensible conclusions about effect sizes, even if the individual study samples are too small to provide either sufficient power or stable results. There is merit to this position, but there are also two problems with this solution to the problem of small samples. First, it creates a two-tiered structure in which the primary researchers do all the work, with little possibility of rewards (i.e., they do studies that cannot be published because of insufficient power and sensitivity) and the meta-analyst gets all the credit for amassing this material into an interpretable whole. Second, it leaves the meta-analyst at the mercy of a pool of primary researchers. Unless there are many studies examining exactly the question the meta-analyst wants to answer, the only alternatives are to change the question or to aggregate together studies that in fact differ in important ways. Neither alternative seems attractive, and if power analysis becomes routine, neither will be strictly necessary. If future studies include large samples and sensitive procedures, then the need for meta-analyses will become less pressing than it is today.

The decision to use large samples is itself likely to improve other aspects of the research. For example, if researchers know that they will have to devote considerable time and resources to data collection, they will probably take more care to pretest, use reliable measures, follow well-laid-out procedures, and so forth. In contrast, if running a study amounts to little more than rounding up 25 undergraduates and herding them to a lab, the need for careful planning, precise measurement and so on. may not be pressing. In large sample research, there may only be one chance to get things right, and researchers are less likely to rely on shoddy measures, incomplete procedures, and so forth. The net result of all this is that studies carried out with careful attention to statistical power are likely to be better and more useful than studies carried out with little regard for power.

Focus on Effect Size. A scan of social science journals confirms that the outcomes of significance tests are routinely reported, but effect size information is often nowhere to be found. Statistical training tends to focus attention on *p* values and significance levels, and not on the substantive question of how well our treatments, interventions, tests, and so on, work (See Cowles, 1989, for a historical analysis of why social scientists focus on significance tests). One of the most important advantages of statistical power analysis is that it makes it virtually impossible to ignore effect sizes.

The whole point of statistical analysis is to help researchers understand data, and it has become increasingly clear over the years that an exclusive focus on significance testing is an impediment to understanding what the data mean (J. Cohen, 1994; Schmidt, 1994; Wilkinson et al., 1999). Statistical power analysis forces them to think about the sort of effect they expect, or at least about the sort of effect they want to be able to detect, and once they start thinking along these lines, it is unlikely they will forget to think about the sort of effect they actually did find. If power analysis did nothing more than direct researchers' attention to the size of their effects, it would be well worth the effort.

Costs Associated With Power Analysis

Statistical power analysis brings a number of benefits, but there are also costs. Most notably, researchers who pay attention to statistical power will find it harder to carry out studies than researchers who do not think about power when planning or evaluating studies. Most researchers (the authors included) have done studies with small samples and insufficient power, and have "gotten away with it," in the sense that they reported significant results. Even when power is low, there is always some chance that you will reject H_0, and a clever researcher can make a career out of "getting lucky." Power analysis will lead researchers to do fewer small sample studies, which in the long run might mean fewer studies period. It is relatively easy to do a dozen small sample studies, with the knowledge that some will and some will not work. It is not so easy to do a dozen large sample studies, and one long-term result of applying power analysis is that the sheer number of studies performed in a field might go down. This is not a bad thing, at least if many low quality, small sample studies are replaced with a few higher quality, large sample studies. Nevertheless, the prospects for building a lengthy vita by doing dozens of studies might be diminished if serious attention is paid to power analysis.

The most serious cost that might be associated with the widespread use of power analysis is an overemphasis on scientific conservatism. If studies are hard to carry out, and require significant resources (time, money, energy), there may be less willingness to try new ideas and approaches, or to test creative hypotheses. The long-term prospects for scientific progress are not good if researchers are unwilling or unable to take risks or try new ideas.

Implications of Power Analysis: Can Power Be Too High?

Throughout this book has advocated paying attention to the probability that researchers will be able to reject a null hypothesis they believe to be wrong (i.e., power). The pitfalls of low power are reasonably obvious, but it is worth considering whether power can be too high. Suppose researchers follow the advice laid out in this book, and design a study with a very high level of power (e.g., power = .95). One implication is there is little real doubt about the outcomes of the statistical tests; with few exceptions, the tests will yield significant outcomes.

When power is extreme (either high or low), researchers are not likely to learn much by conducting a formal hypothesis test. This might imply that power can be too high. This is unlikely. Even when the outcome of a formal statistical hypothesis tests is virtually known in advance, statistical analysis still has clear and obvious value. First, the statistics used in hypothesis testing usually provide an effect size estimate, or the information needed to make this estimate. Even if the statistical test itself provided little new information (with very high or very low power, researchers know how things will turn out), the process of carrying out a statistical test usually provides information that can be used to evaluate the stability and potential replicability of the results.

Consider, for example, the familiar *t* test. The *t* statistic is a ratio of the difference between sample means to the standard error of the difference.[1] If the level of power in the study is very high, then tests of the significance of *t* might not be all that informative; high power means that it will exceed the threshold for significance in virtually all cases. However, the value of the *t* is still informative, because it gives researchers an easy way of determining the standard error term, which in turn can be used in forming confidence intervals. For example, if the $M_1 - M_2 = 10.0$ and $t = 2.50$, then it follows that the standard error of the difference between the means is 4.0 (i.e., 10.0/2.5), and a 95% confidence interval for the difference between means would be 7.84 units wide (i.e., 1.96 * 4.0). This confidence interval gives a very concrete indication of how much variation one might expect from study to study when comparing M_1 to M_2.

[1] The formula for t is $t = \frac{M_1 - M_2}{SE_{M_1} - _{M_2}}$ or $t = \frac{M_1 - M_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

In general, the standard error terms for test statistics tend to become smaller as samples get larger. This is a concrete illustration of the general principle that large samples provide stable and consistent statistical estimates, whereas small samples provide unstable estimates. Even in settings where the significance of a particular statistical test is not in doubt, confidence intervals provide very useful information. Obtaining a confidence interval allows researchers to determine just how much sampling error they might expect in statistical estimates.

The previous paragraph illustrates a distinction that is sometimes blurred by researchers (i.e., the distinction between statistical analysis and null hypothesis testing). Researchers in the behavioral and social sciences have tended to overemphasize formal hypothesis tests, and have paid too little attention to critical questions such as "how large is the effect of the treatments studied here?" (J. Cohen, 1994; Cowles, 1989; Wilkinson et al., 1999). Ironically, serious attention to the topic of power analysis is likely to reduce researchers' dependence on significance testing. The more researchers know about power, the more likely they are to take steps to maximize statistical power, which means that rejecting the null should be nearly a foregone conclusion. Once they understand that the null hypothesis test is *not* the most important facet of your statistical analysis, they are likely to turn attention to the aspects of your analysis that are more important, such as estimating effect sizes.

Does all this mean that null hypothesis tests should be abandoned? Probably not. First, as noted throughout this book, many of the outstanding criticisms of the null hypothesis can be easily addressed by shifting from tests of point hypotheses (e.g., that treatments have no effect whatsoever) to tests of range or interval hypotheses (e.g., that the effects of treatments fall within some range of values denoting small effects). Second, the prospects for fundamental changes in research strategies seem poor, judging from the historical record (Cowles, 1989). Statisticians have been arguing for decades that the use of confidence intervals is preferable to the use of null hypothesis tests, with little apparent effect on actual research practice. Critics of null hypothesis tests have not suggested an alternative that is both viable and that is likely to be widely adopted. There is every indication that null hypothesis tests are here to stay, and careful attention should be given to methods of making the process of hypothesis testing as useful and informative as possible. Careful attention to the principles of power analysis is likely to lead to better research and better statistical analyses.

CONCLUSIONS

Power analysis has profound implications for statistical hypothesis testing, regardless of whether the traditional null hypothesis (i.e., that treatments had no effect whatsoever) or a minimum-effect null hypothesis (e.g., that the effects of treatments are at best small) is tested. In tests of the traditional null, Type I errors are very unlikely (because the traditional null is essentially false by definition), and the only real concern in structuring significance tests should be for maximizing power. This can be done by collecting huge samples, or by using extremely sensitive procedures, but there is an easier way to accomplish this goal. When testing the traditional null hypothesis, always set the alpha level as high as possible. Unfortunately, the weight of tradition rarely allows researchers to choose an alpha level higher than .05 (alpha of .10 is seen in some social science research, but even there, it is barely tolerated). Never choose a more stringent level, unless there is some very unusual and compelling reason to do so.

Tests of minimum-effect null hypotheses are less familiar, but in fact virtually everything known about hypothesis testing applies to tests of this sort. In fact, much of what is already known about hypothesis testing applies better to tests of minimum-effect null hypotheses than to tests of the traditional null.

When testing minimum-effect null hypotheses, seriously consider the possibility that the null hypothesis will be true. This opens the door for what is sometimes considered statistical heresy (i.e., accepting the null). It also opens the door to the possibility that Type I errors will be made, which means that alpha levels, statistical tests, balancing Type I and Type II errors, and so on have some real meaning in this context.

Power analysis has many benefits. It helps researchers make informed decisions about the design of their own research (especially about the number of cases needed and the choice of alpha levels), and it also helps them make sense out of other researchers' significant or nonsignificant results. However, the indirect benefits of power analysis may be the most important. Power analysis is likely to lead them to use larger samples, which in turn will often encourage them to use better measures, more pretests, more carefully designed procedures, and so on. Power analysis also helps to focus attention on effect sizes, rather than focusing exclusively on the *p* value associated with some statistical test. All of this is likely to improve the quality and consistency of social science research.

There are some costs associated with using power analysis. In particular, is often it hard to obtain samples large enough to provide sufficient power. This is especially true in studies where the central hypothesis involves some complex higher order interaction between multiple independent variables. It can be prohibitively difficult to obtain enough power to sensibly test such hypotheses. Reliance on power analysis may also indirectly discourage researchers from trying out new concepts, hypotheses, or procedures.

Throughout, this book has advocated careful attention to statistical power. As the level of power increases, there should be less and less doubt about the outcomes of null hypothesis tests. Ironically, careful attention to power is likely to decrease the relative importance of null hypothesis tests, and increase the attention paid to other aspects of the statistical analysis, notable effect sizes, and confidence intervals. This is a good thing. Null hypothesis testing is valuable (especially when testing minimum-effect null hypotheses), but it should not be the primary focus of statistical analysis. Rather, well-conducted null hypothesis tests should be only a part of the analytic arsenal brought to bear when attempting to determine what data really mean. Power analysis is the first step in carrying out sensible tests of both traditional and minimum-effect null hypotheses.

References

Algina, J., & Keselman, H. J. (1997). Detecting repeated measures effects with univariate and multivariate statistics. *Psychological Methods*, 208–218.

Bunce, D., & West, M. A. (1995). Self perceptions and perceptions of group climate as predictors of individual innovation at work. *Applied Psychology: An International Review*, *44*, 199–215.

Carroll, J. B. (1993). *Human cognitive abilities: A survey offactor-analytic studies*. Cambridge, England: Cambridge University Press.

Cascio, W. F., & Zedeck, S. (1983). Open a new window in rational research planning: Adjust alpha to maximize statistical power. *Personnel Psychology*, *36*, 517–526.

Chow, S.L. (1988). Significance test or effect size? *Psychological Bulletin*, *103*, 105–110.

Clapp, J. F., & Rizk, K. H. (1992). Effect of recreational exercise on midtrimester placental growth. *American Journal of Obstetrics and Gynecology*, *167*, 1518–1521.

Cohen, J. (1962). The statistical power of abnormal-social psychological research. *Journal of Abnormal and Social Psychology*, *65*, 145–153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*, 997–1003.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysisfor the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, *2*, 161–173.

Cotton, M. M., & Evans, K. M. (1990). An evaluation of the Irlen lenses as a treatment for specific reading disorders. *Australian Journal of Psychology*, *42*, 1–12.

Cowles, M. (1989). *Statistics in psychology: An historical perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, *37*, 553–558.

Evans, B. J., & Drasdo, N. (1991). Tinted lenses and related therapies for learning disabilities: A review. *Ophthalmic and Physiological Optics*, *11*, 206–217.

Fick, P. L. (1995). Accepting the null hypothesis. *Memory and Cognition, 23,* 132–138.

Frick, P. J., Silverthorne, P., & Evans, C. (1994). Assessment of childhood anxiety using structured interviews: Patterns of agreement among informants and association with maternal anxiety. *Psychological Assessment, 6,* 372–379.

Gillett, R. (1994). Post hoc power analysis. *Journal of Applied Psychology, 79,* 783–785.

Greenhouse, S. W., & Geisser, S. (1959). On method in the analysis of profile data. *Psychometrika, 24,* 95–112.

Greenwald, A. G. (1993). Consequences of prejudice against the null hypothesis. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 419–448). Hillsdale, NJ: Lawrence Erlbaum Associates.

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology, 79,* 314–316.

Haase, R. F., Waechter, D. M., & Solomon, G. S. (1982). How significant is a significant difference? Average effect size of research in counseling psychology. *Journal of Counseling Psychology, 29,* 58–65.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52,* 15–24.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? *American Psychologist, 42,* 443–455.

Himel, H. N., Liberati, A., Laird, R. D., & Chalmers, T. C. (1986). Adjuvant chemotherapy for breast cancer: A pooled estimate based on published randomized control trials. *Journal of the American Medical Association, 256,* 1148–1159.

Horton, R. L. (1978). *The general linear model: Data analysis in the social and behavioral sciences.* New York: McGraw-Hill.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72–98.

Hunter, J. E., & Hirsh, H. R. (1987). Applications of meta-analysis. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 321–357). Chichester, England: Wiley.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Irlen, H. (1993). *Scotopic sensitivity and reading disability.* Paper presented at the 91st annual convention of the American Psychological Association, Anaheim, CA.

Kraemer, H. C. & Thiemann, S. (1987). *How many subjects?* Newbury Park, CA: Sage.

Labovitz, S. (1968). Criteria for selecting a significance level: A note on the sacredness of .05. *American Sociologist, 3,* 220–222.

Landy, F. J., Farr, J. L., & Jacobs, R. R. (1982). Utility concepts in performance measurement. *Organizational Behavior and Human Performance, 30,* 15–40.

Lipsey, M. W. (1990). *Design sensitivity.* Newbury Park, CA: Sage.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. *American Psychologist, 48,* 1181–1209.

Martin, F., Mackenzie, B., Lovegrove, W., & McNicol, D. (1993). Irlen lenses in the treatment of specific reading disability: An evaluation of outcomes and processes. *Australian Journal of Psychology, 45*, 141–150.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analysing data*. Belmont, CA: Wadsworth.

McDaniel, M. A. (1988). Does pre-employment drug use predict on-the-job suitability? *Personnel Psychology, 41*, 717–729.

Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of psychology. *Journal of Consulting and Clinical Psychology, 46*, 806–834.

Mone, M. A., Mueller, G. C., & Mauland, W. (1996).The perceptions and usage of statistical power in applied psychology and management research. *Personnel Psychology, 49*, 103–120.

Morrison, D. E., & Henkel, R. E. (1970). *The significance test controversy: A reader.* Chicago: Aldine.

Murphy, K. (1990). If the null hypothesis is impossible, why test it? *American Psychologist, 45*, 403–404.

Murphy, K., & Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *Journal of Applied Psychology, 84*, 234–248.

Nagel, S. S., & Neff, M. (1977). Determining an optimal level of statistical significance. *Evaluation Studies Review Annual, 2*, 146–158.

Neale, M. B. (1973). *The Neale Analysis of Reading Ability.* London: Macmillan.

Patnaik, P. B. (1949). The non-central $\div^{2}$- and F-distributions and their applications. *Biometrika, 36*, 202–232.

Pearson, E. S., & Hartley, H. O. (1951). Charts of a power function for analysis of variance tests, derived from the non-central F-distribution. *Biometrika, 38*, 112–130.

Rosenthal, R. (1991). *Meta-analytic procedures for social research.* Newbury Park, CA: Sage.

Rosenthal, R. (1993). Cumulating evidence. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 519–559). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rouanet, H. (1996). Bayesian methods for assessing the importance of effects. *Psychological Bulletin, 119*, 149–158.

Schmidt, F. L. (1992). What do the data really mean? Research findings, meta-analysis and cumulative knowledge in psychology. *American Psychologist, 47*, 1173–1181.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*, 115–129.

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology, 71*, 432–439.

Schmidt, F. L., Mack, M. J., & Hunter, J. E. (1984). Selection utility in the occupation of U.S. park ranger for three modes of test use. *Journal of Applied Psychology, 69*, 490–497.

Schmitt, N., Gooding, R. Z., Noe, R. D., & Kirsch, M. (1984). Metaanalyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37,* 407–422.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies ? *Psychological Bulletin,* 105, 309- 316.

Serlin, R. A., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40,* 73–83.

Serlin, R. A., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbookfor data analysis in the behavioral sciences: Methodological issues* (pp. 199–228). Hillsdale, NJ: Lawrence Erlbaum Associates.

Siegel, S. (1956). *Nonparametric statisticsfor the behavioral sciences.* New York: McGraw-Hill.

Stevens, J. P. (1980). Power of multivariate analysis of variance tests. *Psychological Bulletin, 86,* 728–737.

Stevens, J. (1988). *Applied multivariate statistics for the social sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Stevens, J. P. (2002). *Applied Multivariate Statisticsfor the Social Sciences.* (4th ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tatsuoka, M. (1993a). Effect size. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 461–179). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tatsuoka, M. (1993b). Elements of the general linear model. In G. Keren & C. Lewis (Eds.), *A handbookfor data analysis in the behavioral sciences: Statistical issues* (pp. 3–42). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tiku, M. L., & Yip, D. N. Y. (1978). A four-moment approximation based on the F distribution. *Australian Journal of Statistics, 20* (3), 257–261.

Vonesh, E. F., & Schork, M. A. (1986). Sample sizes in the multivariate analysis of repeated measurements. *Biometrics, 42,* 601–610.

Wilcox, R. R. (1992). Why can methods for comparing means have relatively low power, and what can you do to correct the problem? *Current Directions in Psychological Science, 1,* 101–105.

Wilkinson, L., & Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604

Winter, S. (1987). Irlen lenses: An appraisal. *Australian Educational and Developmental Psychologist, 4,* 1–5.

Yuen, K. K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika, 61,* 165–170.

Zimmerman, D. W., & Zumbo, B. D. (1993). The relative power of parametric and nonparametric statistical methods. In G. Keren & C. Lewis (Eds.), *A handbookfor data analysis in the behavioral sciences: Methodological issues* (pp. 481–518). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zwick, R., & Marascuilo, L. A. (1984). Selection of pairwise comparison procedures for parametric and nonparametric analysis of variance models. *Psychological Bulletin, 95,* 148–155.

Appendix A

Working With the Noncentral *F* Distribution

The noncentral *F* distribution is hard tabulate, because it depends on three separate parameters (df$_{\text{hyp}}$, df_{err} and λ), but there is a reasonably simple approximation based on the central *F* distribution (Horton, 1978; Patnaik, 1949; see also Tiku & Yip, 1978). The noncentral *F* [$F(df_{\text{hyp}}, df_{\text{err}})$] is distributed approximately as the central *F* [$F^*(g, df_{\text{err}})$], with:

$$g = (df_{\text{hyp}} + \lambda)^2 / (df_{\text{hyp}} + 2\lambda) \qquad \text{[A1]}$$

and

$$F^*(g, df_{\text{err}}) = F(df_{\text{hyp},}\, df_{\text{err},}\, p_1) / [(df_{\text{hyp}} + \lambda)/ df_{\text{hyp}}] \qquad \text{[A2]}$$

where $F_{(}df_{\text{hyp},}\, df_{\text{err}}, p_{1)}$ is the central *F* value that has a probability value of p_1. Equation A2 estimates the central *F* that approximates any specific noncentral *F* value.

For example, suppose that in a study where 123 subjects are randomly assigned to one of three treatments, it is found that $SS_{\text{treatments}} = 10$ and $MS_e = 2$ (which means that $R^2 = .04$, or treatments account for 4% of the variance in this study). The *F* statistic used to test the hypothesis that treatments have no effect would have degrees of freedom of $df_{\mathrm{hyp}} = 2$ and $df_{\mathrm{eff}} =$ *120, and the value of F* would be 2.50 [i.e, $\mathrm{MS}_{\mathrm{treatments}} = 10/df_{\mathrm{hyp}} = 10/2 = 5$; $F(2,120) = 5/2 = 2.50$]. The sum of squares for treatments and the mean square error provide enough information to estimate the noncentrality parameter ($\lambda = 10/2 = 5.0$), and this in turn allows researchers to estimate the noncentral F distribution for their study.

Suppose a study intends to test the traditional null hypothesis of no treatment effect. The critical value of F for testing the traditional null with degrees of freedom of 2 and 120 and $\alpha = .05$ is 3.07. Using equations A1 and A2, it is found that $g = 4.08$, and $F^*(4, 120) = .877$. At 4 and 120 degrees of freedom, approximately 50% of the central F distribution lies above a value of .877. Thus, the power of this statistical test is about .50, given the effect found in the study.

Another alternative is to use a computer program specifically designed for computing probabilities in the noncentral F distribution. For example, it is possible to obtain estimates of the appropriate noncentral F using simple functions that are built into a number of widely used programs (e.g., SPSS, Excel). In SPSS, the statement:

$$\text{COMPUTE } p = \text{NCDF.F}(F_{\text{observed}}, df_{\text{hyp}}, df_{\text{err}}, df_{\text{err}} * PV/(1 - PV))$$

calculates the percentage of the noncentral F distribution that is below the value of the observed F [where $\lambda = df_{\mathrm{err}} * PV/(1 - PV)$]. So, if researchers wanted to test the hypothesis that treatments accounted for 1% or less of the variance in outcomes, all they would need to do would be to insert the F obtained in the study into the previous COMPUTE statement (using $PV = .01$ to reflect the minimum effect of interest). Researchers can reject the hypothesis (with $\alpha = .05$) that the treatments account for 1% or less of the variance if the value obtained from this COMPUTE statement is .95 or greater.

Appendix B

One-Stop *F* Table

One Stop F Table

	Hyp F									df_{hyp}								
df_{err}	for	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
3	nil α=.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.62	8.59	8.57	8.55
	nil α=.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.50	26.41	26.32	26.22
	pow .5	8.26	7.21	6.78	6.54	6.39	6.29	6.22	6.16	6.12	6.08	6.02	5.97	5.91	5.86	5.82	5.80	5.76
	pow .8	18.17	15.70	14.83	14.42	14.19	13.93	13.85	13.69	13.66	13.64	13.55	13.48	13.42	13.30	13.26	13.17	12.66
	1% α=.05	10.43	9.70	9.37	9.19	9.07	8.99	8.93	8.88	8.84	8.81	8.77	8.72	8.67	8.63	8.60	8.58	8.55
	1% α=.01	35.15	31.28	29.75	28.93	28.41	28.05	27.79	27.59	27.44	27.31	27.12	26.93	26.73	26.53	26.43	26.33	26.23
	pow .5	8.53	7.33	6.85	6.60	6.44	6.33	6.25	6.19	6.14	6.10	6.04	5.98	5.92	5.86	5.83	5.81	5.79
	pow .8	18.72	16.04	14.96	14.50	14.25	13.98	13.89	13.82	13.69	13.66	13.56	13.48	13.42	13.34	13.31	13.25	13.07
	5% α=.05	11.72	10.30	9.76	9.48	9.30	9.18	9.09	9.02	8.97	8.92	8.86	8.79	8.73	8.66	8.63	8.59	8.56
	5% α=.01	39.41	33.23	31.00	29.84	29.13	28.64	28.30	28.03	27.82	27.66	27.41	27.15	26.90	26.64	26.52	26.39	26.26
	pow .5	9.57	7.82	7.17	6.83	6.62	6.48	6.38	6.30	6.24	6.19	6.12	6.04	5.97	5.89	5.86	5.82	5.80
	pow .8	20.77	17.02	15.60	14.98	14.63	14.30	14.16	14.07	13.90	13.86	13.72	13.62	13.51	13.40	13.38	13.33	13.18
4	nil α=.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5 75	5.72	5.69	5.66
	nil α=.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.84	13.75	13.65	13.56
	pow .5	6.68	5.48	5.00	4.73	4.55	4.43	4.34	4.27	4.22	4.17	4.10	4.03	3.96	3.89	3.86	3.82	3.78
	pow .8	14.17	11.30	10.22	9.66	9.24	9.02	8.86	8.75	8.60	8.53	8.44	8.30	8.19	8.04	7.95	7.87	7.80
	1% α=.05	8.02	7.08	6.68	6.45	6.31	6.20	6.13	6.07	6.03	5.99	5.93	5.87	5.81	5.75	5.72	5.69	5.66
	1% α=.01	22.05	18.36	16.92	16.14	15.65	15.31	15.06	14.87	14.72	14.60	14.42	14.24	14.05	13.86	13.76	13.66	13.56
	pow .5	6.94	5.60	5.07	4.78	4.60	4.46	4.37	4.30	4.24	4.19	4.12	4.05	3.97	3.90	3.86	3.82	3.79
	pow .8	14.64	11.50	10.34	9.75	9.39	9.07	8.91	8.78	8.69	8.56	8.46	8.32	8.20	8.05	7.96	7.88	7.88
	5% α=.05	9.31	7.67	7.05	6.72	6.52	6.38	6.28	6.20	6.14	6.09	6.02	5.94	5.86	5 79	5.75	5.71	5.67
	5% α=.01	25.49	19.86	17.85	16.81	16.17	15.74	15.42	15.19	15.00	14.85	14.63	14.40	14.17	13.93	13.82	13.70	13.58
	pow .5	8.05	6.10	5.39	5.01	4.77	4.61	4.50	4.40	4.33	4.28	4.19	4.10	4.02	3.93	3.88	3.84	3.80
	pow .8	16.63	12.42	10.93	10.18	9.65	9.36	9.15	9.00	8.82	8.73	8.61	8.43	8.25	8.12	8.00	7.91	7.78
5	nil α=.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.73	4.68	4.62	4.56	4.50	4.46	4.43	4.40
	nil α=.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.38	9.29	9.20	9.11
	pow .5	5.91	4.66	4.14	3.87	3.70	3.57	3.48	3.41	3.35	3.31	3.23	3.15	3.08	3.00	2.96	2.91	2.87
	pow .8	12.35	9.38	8.19	7.60	7.24	6.94	6.77	6.59	6.50	6.43	6.28	6.11	5.97	5.83	5.74	5.65	5.54
	1% α=.05	6.94	5.93	5.50	5.26	5.10	4.99	4.91	4.85	4.80	4.76	4.70	4.63	4.57	4.50	4.47	4.44	4.40
	1% α=.01	17.07	13.61	12.26	11.54	11.08	10.76	10.53	10.35	10.21	10.10	9.93	9.75	9.58	9.39	9.30	9.21	9.12
	pow .5	6.20	4.78	4.24	3.93	3.75	3.61	3.51	3.44	3.37	3.33	3.25	3.17	3.09	3.00	2.96	2.92	2.88
	pow .8	12.85	9.59	8.38	7.68	7.30	6.99	6.81	6.68	6.53	6.46	6.30	6.17	5.98	5.84	5.76	5.66	5.54
	5% α=.05	8.31	6.54	5.88	5.53	5.32	5.17	5.06	4.98	4.91	4.86	4.78	4.70	4.62	4.54	4.49	4.45	4.41
	5% α=.01	20.28	14.97	13.10	12.13	11.54	11.14	10.85	10.63	10.45	10.31	10.10	9.89	9.68	9.46	9.35	9.24	9.13
	pow .5	7.42	5.33	4.56	4.18	3.94	3.77	3.65	3.55	3.48	3.42	3.32	3.23	3.13	3.03	2.98	2.93	2.88
	pow .8	14.92	10.51	8.90	8.11	7.64	7.27	7.05	6.84	6.72	6.62	6.44	6.24	6.07	5.89	5.78	5.67	5.55

One Stop F Table

	Hyp F									df_{hyp}							
df_{err}	for	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
6	nil α=.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.81	3.77	3.74
	nil α=.01	13.74	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.23	7.14	7.06
	pow .5	5.45	4.15	3.67	3.39	3.21	3.08	2.99	2.92	2.86	2.82	2.74	2.66	2.58	2.49	2.45	2.40
	pow .8	11.33	8.29	7.15	6.51	6.12	5.84	5.65	5.50	5.38	5.28	5.10	4.95	4.80	4.65	4.56	4.46
	1% α=.05	6.35	5.30	4.85	4.60	4.44	4.33	4.24	4.18	4.13	4.08	4.02	3.95	3.89	3.82	3.78	3.74
	1% α=.01	14.56	11.25	9.98	9.29	8.85	8.55	8.33	8.16	8.03	7.92	7.76	7.59	7.42	7.24	7.15	7.06
	pow .5	5.77	4.32	3.75	3.45	3.25	3.12	3.02	2.95	2.89	2.84	2.76	2.68	2.59	2.50	2.45	2.41
	pow .8	11.86	8.54	7.28	6.60	6.19	5.90	5.69	5.53	5.41	5.31	5.16	5.00	4.82	4.66	4.56	4.46
	5% α=.05	7.82	5.94	5.25	4.89	4.66	4.51	4.40	4.31	4.24	4.19	4.11	4.02	3.94	3.85	3.80	3.76
	5% α=.01	17.73	12.58	10.78	9.86	9.29	8.91	8.63	8.42	8.25	8.12	7.92	7.72	7.51	7.30	7.20	7.09
	pow .5	7.11	4.88	4.13	3.72	3.47	3.29	3.17	3.06	2.98	2.93	2.83	2.74	2.64	2.53	2.48	2.42
	pow .8	14.05	9.45	7.88	7.04	6.53	6.18	5.93	5.70	5.55	5.43	5.26	5.08	4.90	4.69	4.60	4.49
8	nil α=.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.08	3.04	3.00
	nil α=.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.20	5.12	5.03
	pow .5	4.94	3.63	3.12	2.82	2.66	2.52	2.41	2.36	2.30	2.24	2.18	2.11	2.04	1.95	1.91	1.86
	pow .8	10.22	7.17	5.99	5.33	4.95	4.65	4.44	4.30	4.16	4.05	3.91	3.75	3.59	3.41	3.33	3.23
	1% α=.05	5.74	4.64	4.18	3.92	3.75	3.63	3.54	3.47	3.42	3.37	3.31	3.24	3.16	3.09	3.05	3.01
	1% α=.01	12.14	8.99	7.79	7.15	6.74	6.46	6.25	6.09	5.96	5.86	5.70	5.54	5.38	5.21	5.13	5.04
	pow .5	5.36	3.79	3.22	2.88	2.71	2.56	2.45	2.39	2.32	2.29	2.20	2.13	2.05	1.96	1.91	1.86
	pow .8	10.86	7.42	6.14	5.44	5.03	4.72	4.49	4.34	4.20	4.12	3.93	3.77	3.60	3.42	3.34	3.23
	5% α=.05	7.44	5.37	4.62	4.24	3.99	3.83	3.71	3.62	3.55	3.49	3.40	3.31	3.22	3.12	3.07	3.03
	5% α=.01	15.41	10.35	8.61	7.72	7.18	6.81	6.54	6.34	6.19	6.06	5.86	5.67	5.47	5.27	5.17	5.07
	pow .5	6.94	4.48	3.65	3.20	2.92	2.76	2.62	2.54	2.45	2.38	2.30	2.20	2.10	2.00	1.94	1.88
	pow .8	13.34	8.47	6.79	5.90	5.35	5.01	4.74	4.56	4.39	4.25	4.07	3.88	3.69	3.48	3.37	3.26
10	nil α=.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.84	2.77	2.70	2.66	2.62
	nil α=.01	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.25	4.17	4.08
	pow .5	4.68	3.33	2.83	2.53	2.34	2.21	2.11	2.04	1.98	1.93	1.86	1.81	1.74	1.66	1.62	1.58
	pow .8	9.65	6.58	5.40	4.75	4.34	4.05	3.84	3.68	3.55	3.44	3.28	3.14	2.97	2.78	2.69	2.59
	1% α=.05	5.46	4.31	3.83	3.57	3.39	3.27	3.18	3.11	3.05	3.01	2.94	2.86	2.79	2.71	2.67	2.63
	1% α=.01	11.02	7.93	6.77	6.14	5.75	5.48	5.28	5.12	5.00	4.90	4.75	4.59	4.43	4.26	4.18	4.09
	pow .5	5.14	3.56	2.94	2.60	2.40	2.25	2.15	2.07	2.01	1.99	1.91	1.83	1.75	1.67	1.63	1.58
	pow .8	10.37	6.89	5.57	4.87	4.42	4.12	3.90	3.73	3.59	3.51	3.34	3.16	2.98	2.79	2.71	2.60
	5% α=.05	7.39	5.13	4.34	3.93	3.67	3.50	3.37	3.27	3.20	3.13	3.04	2.94	2.85	2.75	2.70	2.64
	5% α=.01	14.48	9.38	7.64	6.75	6.21	5.85	5.58	5.38	5.23	5.10	4.91	4.72	4.52	4.32	4.22	4.12
	pow .5	6.94	4.35	3.43	2.96	2.67	2.48	2.34	2.23	2.15	2.09	1.99	1.89	1.81	1.71	1.65	1.60
	pow .8	13.14	8.06	6.28	5.37	4.81	4.44	4.16	3.95	3.79	3.66	3.46	3.26	3.07	2.85	2.74	2.63

One Stop F Table

	Hyp F									df_{hyp}								
df_{err}	for	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
12	nil α=.05	4.74	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.47	2.43	2.38	2.34
	nil α=.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.70	3.62	3.54	3.45
	pow .5	4.52	3.17	2.63	2.33	2.15	2.02	1.93	1.86	1.80	1.76	1.66	1.61	1.53	1.47	1.43	1.39	1.35
	pow .8	9.30	6.23	5.03	4.38	3.97	3.69	3.48	3.32	3.20	3.09	2.91	2.76	2.58	2.41	2.31	2.21	2.09
	1% α=.05	5.31	4.12	3.63	3.36	3.18	3.06	2.96	2.89	2.83	2.79	2.71	2.64	2.56	2.48	2.43	2.39	2.34
	1% α=.01	10.40	7.34	6.19	5.57	5.19	4.92	4.72	4.57	4.45	4.35	4.20	4.04	3.88	3.72	3.63	3.54	3.45
	pow .5	5.05	3.38	2.75	2.42	2.21	2.07	1.97	1.89	1.83	1.79	1.68	1.62	1.54	1.47	1.43	1.40	1.36
	pow .8	10.12	6.55	5.22	4.51	4.07	3.77	3.54	3.37	3.24	3.13	2.95	2.78	2.60	2.42	2.31	2.21	2.10
	5% α=.05	7.47	5.04	4.20	3.76	3.49	3.31	3.17	3.07	2.99	2.93	2.83	2.73	2.62	2.52	2.46	2.41	2.35
	5% α=.01	14.07	8.88	7.12	6.23	5.68	5.31	5.05	4.85	4.69	4.56	4.37	4.18	3.98	3.78	3.68	3.57	3.47
	pow .5	7.08	4.26	3.30	2.81	2.51	2.32	2.18	2.08	2.00	1.93	1.80	1.72	1.61	1.52	1.47	1.42	1.36
	pow .8	13.18	7.84	6.00	5.07	4.49	4.11	3.83	3.62	3.46	3.33	3.10	2.91	2.69	2.48	2.36	2.24	2.11
14	nil α=.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.31	2.27	2.22	2.18
	nil α=.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.50	3.35	3.27	3.18	3.09
	pow .5	4.41	3.06	2.52	2.23	2.00	1.88	1.79	1.72	1.67	1.59	1.54	1.46	1.39	1.32	1.28	1.26	1.23
	pow .8	9.06	5.99	4.80	4.16	3.72	3.44	3.23	3.07	2.95	2.82	2.67	2.50	2.33	2.14	2.04	1.95	1.83
	1% α=.05	5.24	4.00	3.50	3.22	3.04	2.91	2.82	2.75	2.69	2.64	2.56	2.49	2.40	2.32	2.27	2.23	2.18
	1% α=.01	10.04	6.96	5.82	5.21	4.83	4.56	4.36	4.21	4.09	3.99	3.84	3.69	3.53	3.36	3.28	3.19	3.10
	pow .5	5.01	3.30	2.66	2.32	2.12	1.93	1.83	1.76	1.70	1.66	1.56	1.47	1.41	1.32	1.28	1.26	1.23
	pow .8	9.98	6.35	5.01	4.30	3.85	3.52	3.30	3.13	3.00	2.89	2.71	2.53	2.35	2.15	2.05	1.95	1.84
	5% α=.05	7.62	5.02	4.13	3.66	3.38	3.19	3.05	2.94	2.86	2.79	2.69	2.58	2.47	2.36	2.31	2.25	2.19
	5% α=.01	13.93	8.61	6.82	5.91	5.36	4.98	4.72	4.51	4.35	4.22	4.03	3.83	3.63	3.43	3.33	3.22	3.11
	pow .5	7.25	4.27	3.26	2.76	2.45	2.20	2.06	1.96	1.88	1.78	1.69	1.58	1.48	1.37	1.32	1.29	1.25
	pow .8	13.32	7.76	5.86	4.90	4.32	3.89	3.61	3.40	3.23	3.08	2.88	2.66	2.45	2.22	2.10	1.98	1.85
16	nil α=.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.27	2.19	2.15	2.10	2.06
	nil α=.01	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.10	3.02	2.93	2.84
	pow .5	4.33	2.98	2.39	2.10	1.92	1.80	1.67	1.61	1.56	1.53	1.44	1.36	1.28	1.22	1.18	1.15	1.12
	pow .8	8.88	5.83	4.61	3.97	3.56	3.28	3.05	2.89	2.77	2.67	2.49	2.32	2.14	1.96	1.86	1.76	1.64
	1% α=.05	5.20	3.92	3.41	3.13	2.94	2.81	2.72	2.64	2.58	2.53	2.46	2.38	2.29	2.20	2.16	2.11	2.06
	1% α=.01	9.81	6.71	5.57	4.96	4.58	4.31	4.12	3.97	3.85	3.75	3.60	3.45	3.29	3.12	3.03	2.94	2.85
	pow .5	5.00	3.25	2.60	2.26	2.00	1.86	1.77	1.65	1.60	1.56	1.46	1.38	1.29	1.23	1.18	1.15	1.13
	pow .8	9.90	6.22	4.87	4.15	3.68	3.37	3.15	2.95	2.82	2.71	2.53	2.35	2.16	1.98	1.87	1.76	1.65
	5% α=.05	7.81	5.04	4.10	3.61	3.32	3.12	2.97	2.86	2.77	2.70	2.59	2.48	2.37	2.25	2.19	2.13	2.07
	5% α=.01	13.91	8.47	6.63	5.71	5.15	4.77	4.49	4.29	4.13	4.00	3.80	3.60	3.39	3.19	3.08	2.97	2.86
	pow .5	7.45	4.31	3.26	2.73	2.36	2.16	2.02	1.86	1.79	1.73	1.60	1.49	1.37	1.28	1.22	1.18	1.14
	pow .8	13.52	7.75	5.79	4.80	4.17	3.77	3.48	3.24	3.07	2.94	2.71	2.49	2.26	2.04	1.92	1.80	1.66

One Stop F Table

	Hyp F									df_{hyp}								
df_{err}	for	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
18	nil α=.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.11	2.06	2.02	1.97
	nil α=.01	8.28	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	2.92	2.84	2.75	2.66
	pow .5	4.21	2.87	2.34	2.05	1.87	1.70	1.62	1.56	1.47	1.44	1.35	1.28	1.21	1.13	1.08	1.06	1.04
	pow .8	8.73	5.68	4.49	3.84	3.44	3.13	2.93	2.77	2.62	2.52	2.35	2.19	2.01	1.82	1.71	1.61	1.50
	1% α=.05	5.19	3.87	3.35	3.06	2.87	2.74	2.64	2.57	2.50	2.45	2.38	2.30	2.21	2.12	2.07	2.02	1.97
	1% α=.01	9.67	6.54	5.39	4.78	4.40	4.13	3.94	3.79	3.67	3.57	3.42	3.27	3.11	2.94	2.85	2.76	2.66
	pow .5	5.01	3.21	2.56	2.16	1.95	1.82	1.67	1.60	1.55	1.47	1.38	1.30	1.22	1.14	1.11	1.06	1.04
	pow .8	9.87	6.13	4.76	4.01	3.56	3.26	3.01	2.84	2.71	2.58	2.39	2.22	2.03	1.84	1.73	1.62	1.50
	5% α=.05	8.02	5.09	4.09	3.59	3.28	3.07	2.92	2.80	2.71	2.64	2.52	2.41	2.29	2.17	2.11	2.05	1.98
	5% α=.01	13.99	8.40	6.52	5.58	5.00	4.62	4.34	4.13	3.97	3.83	3.63	3.43	3.22	3.01	2.90	2.79	2.68
	pow .5	7.65	4.37	3.27	2.66	2.34	2.14	1.94	1.83	1.76	1.65	1.52	1.42	1.31	1.20	1.13	1.09	1.06
	pow .8	13.75	7.77	5.76	4.71	4.10	3.69	3.37	3.14	2.98	2.81	2.59	2.37	2.14	1.91	1.77	1.65	1.52
20	nil α=.05	4.34	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.04	1.99	1.95	1.90
	nil α=.01	8.09	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.78	2.69	2.61	2.52
	pow .5	4.17	2.82	2.29	2.00	1.77	1.66	1.58	1.47	1.43	1.35	1.32	1.21	1.14	1.05	1.01	0.99	0.97
	pow .8	8.63	5.58	4.39	3.75	3.32	3.04	2.83	2.65	2.53	2.41	2.26	2.07	1.90	1.70	1.60	1.50	1.38
	1% α=.05	5.20	3.84	3.30	3.01	2.82	2.69	2.59	2.51	2.45	2.39	2.32	2.23	2.14	2.05	2.00	1.95	1.90
	1% α=.01	9.58	6.41	5.26	4.65	4.27	4.00	3.80	3.65	3.53	3.44	3.29	3.13	2.97	2.80	2.71	2.62	2.52
	pow .5	5.03	3.19	2.47	2.13	1.92	1.73	1.63	1.57	1.47	1.43	1.34	1.23	1.15	1.06	1.04	0.99	0.97
	pow .8	9.87	6.07	4.66	3.93	3.48	3.14	2.92	2.75	2.59	2.49	2.31	2.11	1.92	1.72	1.62	1.51	1.39
	5% α=.05	8.20	5.15	4.11	3.58	3.26	3.04	2.88	2.76	2.67	2.59	2.47	2.36	2.23	2.11	2.05	1.98	1.91
	5% α=.01	14.11	8.38	6.46	5.49	4.91	4.51	4.23	4.02	3.85	3.71	3.51	3.30	3.09	2.88	2.77	2.65	2.54
	pow .5	7.95	4.43	3.30	2.67	2.34	2.12	1.92	1.81	1.68	1.63	1.50	1.35	1.24	1.12	1.06	1.02	0.98
	pow .8	14.02	7.82	5.75	4.68	4.05	3.63	3.30	3.08	2.88	2.74	2.51	2.27	2.04	1.79	1.66	1.54	1.40
22	nil α=.05	4.29	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.22	2.15	2.07	1.98	1.94	1.89	1.84
	nil α=.01	7.94	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.67	2.58	2.49	2.40
	pow .5	4.13	2.79	2.25	1.97	1.74	1.62	1.49	1.44	1.35	1.32	1.24	1.18	1.08	1.00	0.97	0.92	0.90
	pow .8	8.55	5.50	4.32	3.67	3.24	2.96	2.73	2.58	2.43	2.33	2.16	2.00	1.81	1.62	1.52	1.40	1.29
	1% α=.05	5.23	3.82	3.27	2.97	2.78	2.64	2.54	2.46	2.40	2.35	2.27	2.18	2.09	2.00	1.95	1.90	1.84
	1% α=.01	9.53	6.32	5.16	4.55	4.16	3.90	3.70	3.55	3.43	3.33	3.18	3.02	2.86	2.69	2.60	2.50	2.41
	pow .5	5.06	3.18	2.45	2.10	1.89	1.70	1.61	1.49	1.44	1.36	1.27	1.20	1.09	1.01	0.97	0.94	0.90
	pow .8	9.88	6.03	4.60	3.87	3.41	3.08	2.85	2.66	2.52	2.39	2.21	2.04	1.83	1.63	1.53	1.42	1.29
	5% α=.05	8.41	5.22	4.13	3.58	3.25	3.02	2.86	2.73	2.63	2.56	2.44	2.31	2.19	2.06	1.99	1.92	1.85
	5% α=.01	14.26	8.40	6.43	5.44	4.84	4.44	4.15	3.93	3.76	3.62	3.41	3.20	2.99	2.77	2.66	2.54	2.43
	pow .5	8.17	4.51	3.33	2.68	2.34	2.06	1.91	1.75	1.67	1.56	1.43	1.33	1.19	1.07	1.02	0.95	0.92
	pow .8	14.27	7.89	5.77	4.66	4.02	3.56	3.26	3.00	2.82	2.66	2.43	2.20	1.95	1.71	1.58	1.45	1.31

One Stop F Table

	Hyp F									df_{hyp}								
df_{err}	for	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
24	nil α=.05	4.25	3.40	3.01	2.78	2.62	2.51	2.42	2.35	2.30	2.25	2.18	2.11	2.03	1.94	1.89	1.84	1.79
	nil α=.01	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.58	2.49	2.40	2.31
	pow .5	4.10	2.76	2.22	1.94	1.71	1.60	1.47	1.41	1.32	1.30	1.22	1.12	1.02	0.96	0.90	0.87	0.85
	pow .8	8.48	5.44	4.25	3.61	3.18	2.90	2.67	2.52	2.37	2.27	2.10	1.92	1.73	1.54	1.44	1.33	1.21
	1% α=.05	5.26	3.80	3.25	2.94	2.75	2.61	2.50	2.42	2.36	2.31	2.23	2.14	2.05	1.95	1.90	1.85	1.79
	1% α=.01	9.51	6.25	5.08	4.47	4.08	3.81	3.62	3.46	3.34	3.24	3.09	2.93	2.77	2.60	2.51	2.41	2.31
	pow .5	5.10	3.18	2.43	2.08	1.81	1.68	1.53	1.47	1.42	1.34	1.25	1.14	1.08	0.96	0.93	0.89	0.85
	pow .8	9.91	6.00	4.56	3.82	3.33	3.02	2.77	2.60	2.47	2.34	2.15	1.96	1.78	1.56	1.46	1.34	1.22
	5% α=.05	8.63	5.30	4.16	3.59	3.25	3.01	2.84	2.71	2.61	2.53	2.41	2.28	2.15	2.02	1.95	1.88	1.81
	5% α=.01	14.43	8.43	6.42	5.41	4.80	4.39	4.09	3.87	3.69	3.55	3.34	3.13	2.91	2.68	2.57	2.45	2.33
	pow .5	8.38	4.58	3.37	2.69	2.35	2.06	1.91	1.74	1.66	1.55	1.42	1.32	1.17	1.03	0.98	0.92	0.87
	pow .8	14.52	7.97	5.79	4.66	4.01	3.54	3.23	2.96	2.79	2.62	2.38	2.16	1.90	1.64	1.52	1.38	1.24
26	nil α=.05	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.26	2.22	2.15	2.07	1.99	1.90	1.85	1.80	1.75
	nil α=.01	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.50	2.42	2.33	2.23
	pow .5	4.07	2.73	2.20	1.85	1.69	1.52	1.44	1.34	1.30	1.22	1.15	1.10	1.01	0.91	0.86	0.84	0.80
	pow .8	8.42	5.38	4.20	3.53	3.13	2.82	2.62	2.44	2.32	2.20	2.03	1.87	1.68	1.48	1.37	1.27	1.15
	1% α=.05	5.29	3.80	3.23	2.92	2.72	2.58	2.48	2.40	2.33	2.28	2.19	2.11	2.01	1.92	1.86	1.81	1.75
	1% α=.01	9.51	6.21	5.03	4.40	4.01	3.75	3.55	3.39	3.27	3.17	3.02	2.86	2.70	2.52	2.43	2.34	2.24
	pow .5	5.14	3.18	2.42	2.07	1.80	1.66	1.51	1.45	1.35	1.32	1.23	1.12	1.02	0.92	0.90	0.84	0.81
	pow .8	9.95	5.98	4.53	3.78	3.29	2.98	2.73	2.55	2.40	2.29	2.11	1.91	1.71	1.50	1.40	1.28	1.15
	5% α=.05	8.85	5.38	4.20	3.61	3.25	3.01	2.84	2.70	2.60	2.51	2.39	2.26	2.12	1.99	1.92	1.84	1.77
	5% α=.01	14.63	8.48	6.43	5.39	4.77	4.35	4.05	3.82	3.64	3.50	3.29	3.07	2.84	2.62	2.50	2.38	2.26
	pow .5	8.59	4.66	3.41	2.71	2.36	2.06	1.91	1.74	1.66	1.54	1.41	1.26	1.12	1.02	0.95	0.88	0.83
	pow .8	14.78	8.06	5.83	4.67	4.00	3.52	3.21	2.94	2.76	2.58	2.35	2.09	1.84	1.60	1.46	1.32	1.17
28	nil α=.05	4.19	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.23	2.19	2.12	2.04	1.96	1.87	1.82	1.77	1.71
	nil α=.01	7.63	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.44	2.35	2.26	2.17
	pow .5	4.05	2.71	2.18	1.83	1.67	1.50	1.42	1.32	1.28	1.21	1.13	1.04	0.95	0.90	0.83	0.79	0.77
	pow .8	8.38	5.34	4.15	3.49	3.08	2.78	2.58	2.40	2.28	2.16	1.99	1.81	1.62	1.44	1.32	1.21	1.09
	1% α=.05	5.33	3.80	3.22	2.90	2.70	2.56	2.45	2.37	2.30	2.25	2.17	2.08	1.98	1.89	1.83	1.78	1.72
	1% α=.01	9.52	6.17	4.98	4.35	3.96	3.69	3.49	3.34	3.22	3.12	2.96	2.80	2.64	2.46	2.37	2.27	2.17
	pow .5	5.19	3.19	2.42	2.06	1.78	1.65	1.50	1.43	1.33	1.30	1.17	1.11	1.01	0.91	0.86	0.82	0.77
	pow .8	10.00	5.97	4.50	3.75	3.26	2.94	2.69	2.52	2.36	2.25	2.04	1.87	1.67	1.46	1.34	1.23	1.10
	5% α=.05	9.07	5.47	4.25	3.64	3.26	3.01	2.83	2.70	2.59	2.50	2.37	2.24	2.10	1.96	1.89	1.81	1.73
	5% α=.01	14.83	8.55	6.45	5.39	4.75	4.33	4.02	3.79	3.61	3.46	3.24	3.02	2.79	2.56	2.44	2.32	2.19
	pow .5	8.80	4.74	3.45	2.74	2.37	2.07	1.91	1.74	1.59	1.54	1.41	1.25	1.11	0.98	0.91	0.85	0.78
	pow .8	15.04	8.16	5.87	4.69	4.00	3.52	3.19	2.92	2.71	2.56	2.32	2.06	1.81	1.54	1.41	1.27	1.12

One Stop F Table

	Hyp F									df_{hyp}								
df_{err}		for	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
30	nil	α=.05	4.16	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.84	1.79	1.74	1.68
	nil	α=.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.39	2.30	2.21	2.11
		pow .5	4.03	2.69	2.16	1.82	1.65	1.48	1.41	1.30	1.27	1.19	1.12	1.02	0.94	0.86	0.82	0.76	0.73
		pow .8	8.33	5.30	4.12	3.45	3.05	2.74	2.54	2.36	2.24	2.12	1.95	1.77	1.58	1.39	1.28	1.17	1.04
	1%	α=.05	5.38	3.80	3.21	2.89	2.68	2.54	2.43	2.35	2.28	2.23	2.14	2.05	1.96	1.86	1.80	1.75	1.69
	1%	α=.01	9.54	6.15	4.94	4.31	3.92	3.64	3.44	3.29	3.17	3.07	2.91	2.75	2.59	2.41	2.32	2.22	2.12
		pow .5	5.23	3.19	2.41	2.05	1.77	1.64	1.49	1.42	1.32	1.24	1.15	1.05	0.96	0.87	0.82	0.77	0.73
		pow .8	10.06	5.97	4.49	3.73	3.23	2.91	2.66	2.49	2.32	2.19	2.01	1.81	1.61	1.41	1.30	1.17	1.05
	5%	α=.05	9.29	5.57	4.29	3.66	3.28	3.02	2.83	2.69	2.58	2.49	2.36	2.22	2.08	1.94	1.86	1.78	1.70
	5%	α=.01	15.04	8.62	6.47	5.40	4.75	4.31	4.00	3.76	3.58	3.43	3.20	2.98	2.75	2.51	2.39	2.27	2.14
		pow .5	9.01	4.82	3.50	2.76	2.39	2.08	1.92	1.74	1.59	1.54	1.40	1.25	1.11	0.94	0.88	0.82	0.76
		pow .8	15.31	8.26	5.91	4.71	4.01	3.51	3.19	2.91	2.69	2.54	2.30	2.04	1.78	1.50	1.36	1.23	1.07
40	nil	α=.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.74	1.69	1.64	1.58
	nil	α=.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.20	2.11	2.02	1.92
		pow .5	3.97	2.63	2.02	1.76	1.52	1.42	1.29	1.25	1.16	1.08	1.02	0.93	0.86	0.76	0.70	0.65	0.61
		pow .8	8.20	5.16	3.96	3.32	2.89	2.61	2.39	2.23	2.09	1.97	1.80	1.62	1.44	1.23	1.12	1.00	0.88
	1%	α=.05	5.64	3.85	3.21	2.86	2.64	2.49	2.38	2.29	2.22	2.16	2.07	1.97	1.87	1.77	1.71	1.65	1.58
	1%	α=.01	9.75	6.12	4.86	4.20	3.79	3.51	3.30	3.14	3.01	2.91	2.75	2.59	2.42	2.23	2.14	2.03	1.92
		pow .5	5.49	3.26	2.42	2.04	1.75	1.54	1.45	1.33	1.22	1.20	1.12	1.01	0.88	0.77	0.71	0.66	0.61
		pow .8	10.38	6.01	4.46	3.67	3.15	2.80	2.56	2.36	2.19	2.08	1.90	1.70	1.48	1.25	1.14	1.01	0.88
	5%	α=.05	10.44	6.01	4.56	3.83	3.39	3.09	2.88	2.72	2.59	2.49	2.34	2.19	2.03	1.86	1.78	1.69	1.60
	5%	α=.01	16.14	9.06	6.70	5.51	4.80	4.32	3.98	3.72	3.52	3.35	3.11	2.86	2.61	2.36	2.22	2.09	1.95
		pow .5	10.00	5.35	3.73	3.00	2.49	2.15	1.97	1.77	1.61	1.55	1.35	1.19	1.05	0.88	0.80	0.72	0.63
		pow .8	16.64	8.82	6.19	4.91	4.10	3.56	3.20	2.90	2.67	2.51	2.22	1.94	1.67	1.38	1.23	1.07	0.91
50	nil	α=.05	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.95	1.87	1.78	1.68	1.63	1.57	1.51
	nil	α=.01	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.56	2.42	2.26	2.10	2.01	1.91	1.80
		pow .5	3.93	2.59	1.99	1.72	1.49	1.39	1.26	1.15	1.13	1.06	0.99	0.91	0.79	0.70	0.63	0.60	0.54
		pow .8	8.11	5.09	3.88	3.25	2.82	2.54	2.31	2.13	2.01	1.89	1.73	1.55	1.35	1.14	1.02	0.91	0.78
	1%	α=.05	5.93	3.94	3.24	2.87	2.63	2.47	2.35	2.26	2.19	2.12	2.03	1.93	1.83	1.71	1.65	1.59	1.52
	1%	α=.01	10.04	6.18	4.85	4.16	3.74	3.44	3.23	3.07	2.94	2.83	2.67	2.50	2.32	2.13	2.03	1.92	1.81
		pow .5	5.76	3.35	2.46	1.97	1.75	1.54	1.44	1.32	1.21	1.18	1.04	0.94	0.81	0.71	0.67	0.60	0.54
		pow .8	10.74	6.11	4.48	3.63	3.13	2.76	2.52	2.31	2.14	2.03	1.82	1.61	1.39	1.17	1.05	0.92	0.78
	5%	α=.05	11.39	6.50	4.84	4.02	3.53	3.20	2.96	2.78	2.64	2.53	2.36	2.19	2.01	1.83	1.74	1.64	1.54
	5%	α=.01	17.26	9.56	6.98	5.69	4.92	4.40	4.03	3.75	3.53	3.35	3.09	2.82	2.55	2.28	2.14	1.99	1.84
		pow .5	11.09	5.76	4.08	3.16	2.61	2.23	2.03	1.82	1.65	1.58	1.37	1.20	1.00	0.84	0.76	0.66	0.57
		pow .8	17.86	9.36	6.55	5.11	4.24	3.66	3.27	2.95	2.70	2.53	2.22	1.92	1.61	1.31	1.16	0.99	0.81

One Stop F Table

	Hyp F									df_{hyp}								
df_{err}	for	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
60	nil α=.05	3.99	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.83	1.75	1.65	1.59	1.53	1.47
	nil α=.01	7.07	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.03	1.94	1.84	1.73
	pow .5	3.90	2.57	1.97	1.70	1.47	1.30	1.24	1.13	1.04	1.04	0.91	0.84	0.78	0.65	0.58	0.54	0.48
	pow .8	8.06	5.04	3.83	3.20	2.77	2.46	2.26	2.08	1.94	1.85	1.66	1.48	1.30	1.08	0.96	0.84	0.70
	1% α=.05	6.24	4.04	3.29	2.90	2.64	2.47	2.35	2.25	2.17	2.11	2.01	1.91	1.80	1.68	1.62	1.55	1.47
	1% α=.01	10.35	6.28	4.88	4.16	3.72	3.42	3.20	3.03	2.90	2.79	2.62	2.45	2.26	2.07	1.96	1.85	1.73
	pow .5	6.04	3.44	2.50	1.99	1.76	1.54	1.37	1.31	1.20	1.11	1.03	0.93	0.80	0.70	0.63	0.55	0.50
	pow .8	11.13	6.22	4.53	3.65	3.13	2.75	2.47	2.29	2.11	1.97	1.78	1.58	1.35	1.13	0.99	0.86	0.72
	5% α=.05	12.49	6.94	5.14	4.23	3.68	3.31	3.05	2.86	2.70	2.58	2.39	2.20	2.01	1.82	1.72	1.61	1.50
	5% α=.01	18.38	10.06	7.29	5.90	5.07	4.51	4.11	3.81	3.58	3.39	3.11	2.82	2.53	2.24	2.09	1.93	1.77
	pow .5	11.97	6.30	4.33	3.33	2.82	2.41	2.11	1.88	1.77	1.62	1.39	1.21	1.01	0.84	0.72	0.61	0.53
	pow .8	19.10	9.93	6.86	5.33	4.44	3.81	3.36	3.02	2.78	2.57	2.24	1.93	1.60	1.29	1.11	0.93	0.75
70	nil α=.05	3.97	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.89	1.81	1.72	1.62	1.56	1.50	1.43
	nil α=.01	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.58	2.45	2.30	2.15	1.98	1.89	1.78	1.67
	pow .5	3.88	2.55	1.95	1.68	1.46	1.28	1.23	1.12	1.03	1.02	0.90	0.82	0.72	0.60	0.58	0.51	0.45
	pow .8	8.02	5.00	3.80	3.16	2.73	2.43	2.23	2.05	1.91	1.81	1.62	1.44	1.24	1.03	0.92	0.80	0.66
	1% α=.05	6.57	4.14	3.35	2.92	2.66	2.48	2.35	2.25	2.17	2.10	2.00	1.89	1.78	1.66	1.59	1.52	1.44
	1% α=.01	10.67	6.39	4.93	4.18	3.73	3.41	3.19	3.01	2.87	2.76	2.59	2.41	2.23	2.03	1.92	1.80	1.68
	pow .5	6.32	3.54	2.55	2.11	1.78	1.55	1.37	1.32	1.20	1.11	1.03	0.93	0.80	0.66	0.59	0.52	0.45
	pow .8	11.55	6.35	4.59	3.71	3.14	2.75	2.47	2.28	2.10	1.96	1.76	1.55	1.32	1.08	0.95	0.81	0.66
	5% α=.05	13.34	7.42	5.45	4.43	3.84	3.44	3.16	2.94	2.77	2.64	2.44	2.23	2.03	1.81	1.71	1.59	1.48
	5% α=.01	19.46	10.58	7.61	6.13	5.23	4.64	4.21	3.89	3.64	3.44	3.14	2.84	2.53	2.22	2.06	1.89	1.72
	pow .5	13.03	6.69	4.56	3.60	2.94	2.50	2.18	2.02	1.82	1.66	1.49	1.23	1.02	0.80	0.72	0.61	0.50
	pow .8	20.22	10.46	7.20	5.59	4.60	3.93	3.45	3.13	2.84	2.62	2.30	1.94	1.60	1.26	1.09	0.90	0.71
80	nil α=.05	3.95	3.11	2.72	2.49	2.33	2.21	2.12	2.05	2.00	1.95	1.87	1.79	1.70	1.60	1.54	1.48	1.41
	nil α=.01	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.41	2.27	2.11	1.94	1.85	1.75	1.63
	pow .5	3.87	2.54	1.94	1.67	1.44	1.27	1.21	1.11	1.02	0.94	0.89	0.82	0.71	0.60	0.54	0.48	0.42
	pow .8	7.99	4.97	3.77	3.14	2.71	2.40	2.20	2.02	1.88	1.76	1.60	1.42	1.22	1.00	0.88	0.76	0.62
	1% α=.05	6.83	4.26	3.41	2.96	2.69	2.50	2.36	2.26	2.17	2.10	2.00	1.89	1.77	1.64	1.57	1.50	1.42
	1% α=.01	10.98	6.51	4.99	4.22	3.74	3.42	3.19	3.01	2.86	2.75	2.57	2.39	2.20	1.99	1.89	1.77	1.64
	pow .5	6.73	3.64	2.60	2.14	1.80	1.56	1.38	1.32	1.21	1.11	1.03	0.87	0.79	0.65	0.58	0.52	0.44
	pow .8	11.95	6.48	4.66	3.75	3.16	2.76	2.47	2.28	2.10	1.95	1.75	1.52	1.31	1.06	0.93	0.79	0.63
	5% α=.05	14.39	7.84	5.71	4.65	4.01	3.58	3.26	3.03	2.85	2.71	2.49	2.27	2.04	1.82	1.70	1.58	1.46
	5% α=.01	20.52	11.08	7.93	6.35	5.41	4.77	4.32	3.98	3.72	3.50	3.18	2.86	2.54	2.21	2.04	1.87	1.69
	pow .5	13.83	7.22	4.92	3.76	3.06	2.59	2.34	2.08	1.87	1.70	1.52	1.25	1.03	0.80	0.69	0.58	0.47
	pow .8	21.36	11.02	7.55	5.81	4.76	4.06	3.59	3.21	2.91	2.67	2.34	1.97	1.61	1.25	1.06	0.88	0.67

One Stop F Table

	Hyp F									df_{hyp}								
df_{err}	for	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
90	nil α=.05	3.94	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.98	1.94	1.86	1.78	1.69	1.58	1.53	1.46	1.39
	nil α=.01	6.92	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52	2.39	2.24	2.09	1.91	1.82	1.72	1.60
	pow .5	3.86	2.53	1.93	1.66	1.43	1.26	1.21	1.10	1.01	0.94	0.89	0.81	0.70	0.59	0.54	0.46	0.40
	pow .8	7.97	4.95	3.75	3.12	2.69	2.38	2.18	2.00	1.86	1.74	1.58	1.40	1.20	0.98	0.86	0.73	0.59
	1% α=.05	6.97	4.37	3.48	3.00	2.71	2.52	2.38	2.26	2.18	2.11	2.00	1.88	1.76	1.63	1.56	1.48	1.40
	1% α=.01	11.29	6.64	5.06	4.26	3.77	3.43	3.19	3.01	2.86	2.74	2.56	2.38	2.18	1.97	1.86	1.74	1.61
	pow .5	6.86	3.74	2.66	2.18	1.83	1.58	1.47	1.33	1.21	1.12	1.03	0.87	0.79	0.65	0.55	0.49	0.40
	pow .8	12.12	6.62	4.74	3.79	3.19	2.78	2.52	2.28	2.10	1.95	1.75	1.51	1.29	1.04	0.89	0.76	0.60
	5% α=.05	15.17	8.31	6.02	4.88	4.15	3.70	3.37	3.12	2.93	2.77	2.54	2.30	2.07	1.83	1.70	1.58	1.44
	5% α=.01	21.57	11.59	8.25	6.59	5.58	4.92	4.44	4.08	3.80	3.57	3.24	2.90	2.55	2.21	2.03	1.85	1.66
	pow .5	14.87	7.58	5.14	3.92	3.29	2.78	2.41	2.14	1.92	1.82	1.55	1.34	1.09	0.85	0.69	0.58	0.45
	pow .8	22.43	11.51	7.87	6.04	4.97	4.23	3.70	3.30	2.99	2.77	2.39	2.03	1.65	1.27	1.06	0.87	0.65
100	nil α=.05	3.93	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.85	1.77	1.67	1.57	1.51	1.45	1.37
	nil α=.01	6.89	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.37	2.22	2.07	1.89	1.80	1.69	1.57
	pow .5	3.85	2.52	1.92	1.66	1.43	1.26	1.20	1.10	1.01	0.93	0.88	0.80	0.70	0.59	0.50	0.46	0.39
	pow .8	7.95	4.94	3.73	3.10	2.67	2.37	2.17	1.99	1.84	1.72	1.56	1.38	1.18	0.97	0.83	0.71	0.57
	1% α=.05	7.24	4.49	3.55	3.04	2.74	2.54	2.39	2.28	2.19	2.11	2.00	1.88	1.76	1.62	1.55	1.47	1.38
	1% α=.01	11.60	6.76	5.13	4.30	3.80	3.45	3.21	3.02	2.87	2.75	2.56	2.37	2.17	1.96	1.84	1.72	1.58
	pow .5	7.11	3.84	2.71	2.22	1.85	1.59	1.49	1.34	1.22	1.12	1.04	0.87	0.74	0.61	0.55	0.49	0.39
	pow .8	12.45	6.76	4.82	3.83	3.22	2.80	2.53	2.29	2.11	1.95	1.75	1.50	1.26	1.01	0.88	0.74	0.58
	5% α=.05	16.18	8.81	6.27	5.05	4.32	3.83	3.49	3.21	3.00	2.84	2.60	2.34	2.09	1.84	1.71	1.57	1.43
	5% α=.01	22.59	12.08	8.57	6.82	5.76	5.06	4.56	4.18	3.88	3.65	3.29	2.94	2.58	2.21	2.03	1.84	1.64
	pow .5	15.62	7.93	5.51	4.19	3.40	2.87	2.49	2.29	2.06	1.87	1.58	1.36	1.11	0.86	0.70	0.59	0.44
	pow .8	23.49	12.03	8.22	6.30	5.14	4.36	3.81	3.43	3.10	2.83	2.43	2.06	1.67	1.28	1.06	0.86	0.63
120	nil α=.05	3.91	3.07	2.68	2.45	2.29	2.17	2.09	2.01	1.96	1.91	1.83	1.75	1.66	1.55	1.49	1.43	1.35
	nil α=.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.86	1.76	1.65	1.53
	pow .5	3.84	2.51	1.91	1.56	1.42	1.25	1.11	1.09	1.00	0.92	0.87	0.74	0.64	0.54	0.50	0.43	0.36
	pow .8	7.93	4.91	3.71	3.05	2.65	2.34	2.12	1.97	1.82	1.70	1.54	1.34	1.14	0.92	0.81	0.68	0.53
	1% α=.05	7.76	4.74	3.66	3.13	2.81	2.59	2.43	2.31	2.21	2.13	2.01	1.89	1.75	1.61	1.54	1.45	1.36
	1% α=.01	12.20	7.02	5.28	4.40	3.86	3.50	3.24	3.04	2.88	2.76	2.56	2.36	2.15	1.93	1.81	1.69	1.55
	pow .5	7.58	4.04	2.92	2.29	1.90	1.63	1.52	1.36	1.24	1.13	1.05	0.87	0.74	0.61	0.54	0.46	0.37
	pow .8	13.10	7.05	4.98	3.93	3.29	2.85	2.56	2.32	2.12	1.97	1.75	1.50	1.25	1.00	0.86	0.71	0.55
	5% α=.05	17.88	9.64	6.89	5.45	4.64	4.09	3.70	3.41	3.17	2.98	2.71	2.43	2.15	1.87	1.72	1.57	1.42
	5% α=.01	24.59	13.05	9.20	7.28	6.12	5.35	4.80	4.38	4.06	3.80	3.41	3.02	2.63	2.23	2.03	1.83	1.61
	pow .5	17.37	8.79	5.92	4.63	3.74	3.15	2.73	2.41	2.25	2.04	1.72	1.47	1.13	0.87	0.74	0.59	0.42
	pow .8	25.54	13.02	8.83	6.78	5.51	4.67	4.06	3.61	3.30	3.00	2.57	2.16	1.71	1.29	1.08	0.85	0.61

One Stop F Table

	Hyp F									df_{hyp}								
df_{err}	for	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
150	nil α=.05	3.89	3.06	2.67	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.81	1.73	1.64	1.53	1.47	1.40	1.32
	nil α=.01	6.80	4.75	3.92	3.45	3.14	2.92	2.76	2.63	2.53	2.44	2.30	2.16	2.00	1.83	1.73	1.62	1.49
	pow .5	3.83	2.50	1.90	1.55	1.41	1.24	1.10	1.08	0.99	0.92	0.86	0.73	0.63	0.54	0.45	0.40	0.33
	pow .8	7.90	4.89	3.69	3.02	2.63	2.32	2.09	1.94	1.80	1.68	1.52	1.31	1.11	0.90	0.77	0.64	0.49
	1% α=.05	8.61	5.01	3.86	3.28	2.92	2.66	2.49	2.36	2.25	2.17	2.03	1.90	1.76	1.61	1.53	1.44	1.34
	1% α=.01	13.04	7.40	5.51	4.56	3.98	3.59	3.31	3.09	2.93	2.79	2.58	2.37	2.15	1.92	1.79	1.66	1.51
	pow .5	8.26	4.42	3.09	2.40	1.98	1.78	1.56	1.40	1.27	1.15	1.06	0.88	0.74	0.61	0.51	0.43	0.34
	pow .8	14.11	7.43	5.21	4.09	3.40	2.96	2.62	2.37	2.16	2.00	1.77	1.51	1.25	0.98	0.83	0.68	0.51
	5% α=.05	20.52	10.86	7.64	6.06	5.11	4.48	4.03	3.69	3.41	3.20	2.88	2.57	2.24	1.92	1.75	1.59	1.41
	5% α=.01	27.47	14.46	10.12	7.95	6.65	5.78	5.16	4.69	4.33	4.04	3.60	3.17	2.73	2.28	2.06	1.83	1.59
	pow .5	19.73	10.24	6.86	5.19	4.19	3.52	3.04	2.67	2.48	2.25	1.90	1.61	1.23	0.93	0.75	0.59	0.41
	pow .8	28.49	14.57	9.81	7.46	6.05	5.10	4.43	3.92	3.56	3.24	2.76	2.31	1.81	1.35	1.09	0.85	0.58
200	nil α=.05	3.88	3.04	2.65	2.42	2.26	2.14	2.05	1.98	1.93	1.88	1.80	1.71	1.62	1.51	1.45	1.38	1.30
	nil α=.01	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.27	2.13	1.97	1.79	1.69	1.58	1.45
	pow .5	3.82	2.48	1.89	1.54	1.40	1.23	1.09	1.07	0.98	0.91	0.79	0.72	0.63	0.49	0.45	0.37	0.30
	pow .8	7.88	4.86	3.66	3.00	2.60	2.30	2.07	1.92	1.78	1.65	1.47	1.29	1.09	0.86	0.75	0.60	0.45
	1% α=.05	9.58	5.57	4.22	3.49	3.08	2.81	2.61	2.46	2.33	2.23	2.09	1.93	1.78	1.61	1.52	1.43	1.32
	1% α=.01	14.39	8.02	5.90	4.83	4.18	3.75	3.43	3.20	3.01	2.86	2.63	2.40	2.16	1.91	1.78	1.64	1.48
	pow .5	9.37	4.88	3.36	2.68	2.20	1.88	1.64	1.46	1.40	1.28	1.09	0.96	0.75	0.61	0.51	0.43	0.33
	pow .8	15.40	8.08	5.62	4.38	3.62	3.11	2.74	2.46	2.27	2.09	1.81	1.56	1.26	0.98	0.82	0.66	0.48
	5% α=.05	24.55	12.87	8.94	7.02	5.88	5.11	4.56	4.15	3.84	3.56	3.18	2.79	2.40	2.01	1.82	1.62	1.41
	5% α=.01	32.04	16.72	11.60	9.04	7.51	6.49	5.76	5.21	4.78	4.44	3.93	3.42	2.90	2.38	2.12	1.85	1.58
	pow .5	23.65	11.85	8.17	6.17	4.96	4.16	3.58	3.15	2.81	2.53	2.21	1.79	1.42	1.01	0.81	0.63	0.41
	pow .8	33.10	16.67	11.31	8.57	6.92	5.82	5.04	4.45	3.99	3.61	3.10	2.54	2.00	1.44	1.15	0.88	0.58
300	nil α=.05	3.86	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.78	1.70	1.60	1.49	1.43	1.36	1.27
	nil α=.01	6.72	4.68	3.85	3.38	3.08	2.86	2.70	2.57	2.47	2.38	2.24	2.10	1.94	1.76	1.66	1.55	1.41
	pow .5	3.80	2.47	1.88	1.53	1.39	1.22	1.09	0.98	0.97	0.90	0.78	0.72	0.62	0.48	0.41	0.37	0.26
	pow .8	7.85	4.84	3.64	2.98	2.58	2.28	2.05	1.87	1.75	1.63	1.44	1.27	1.07	0.83	0.71	0.58	0.41
	1% α=.05	11.62	6.54	4.85	3.97	3.43	3.08	2.84	2.65	2.51	2.38	2.20	2.02	1.83	1.64	1.54	1.43	1.31
	1% α=.01	16.85	9.18	6.63	5.36	4.59	4.07	3.70	3.42	3.20	3.03	2.76	2.49	2.22	1.93	1.78	1.62	1.45
	pow .5	11.36	5.83	3.97	3.17	2.56	2.17	1.89	1.67	1.50	1.45	1.23	1.01	0.84	0.62	0.51	0.43	0.30
	pow .8	17.91	9.26	6.37	4.96	4.04	3.44	3.02	2.69	2.44	2.27	1.95	1.63	1.33	1.00	0.82	0.66	0.45
	5% α=.05	32.04	16.65	11.52	8.94	7.35	6.32	5.59	5.05	4.62	4.28	3.77	3.25	2.74	2.22	1.97	1.70	1.44
	5% α=.01	40.62	20.94	14.40	11.13	9.16	7.86	6.92	6.21	5.67	5.23	4.58	3.92	3.26	2.60	2.27	1.94	1.59
	pow .5	31.22	15.66	10.47	7.87	6.49	5.42	4.66	4.09	3.64	3.29	2.75	2.29	1.73	1.22	0.96	0.70	0.43
	pow .8	41.71	20.99	14.09	10.62	8.63	7.22	6.23	5.48	4.90	4.43	3.73	3.07	2.35	1.66	1.31	0.96	0.59

One Stop F Table

	Hyp F									df_{hyp}								
df_{err}	for	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
400	nil α=.05	3.85	3.02	2.63	2.39	2.24	2.12	2.03	1.96	1.90	1.85	1.77	1.69	1.59	1.48	1.42	1.35	1.26
	nil α=.01	6.70	4.66	3.83	3.37	3.06	2.85	2.68	2.56	2.45	2.36	2.23	2.08	1.92	1.74	1.64	1.53	1.39
	pow .5	3.80	2.47	1.88	1.52	1.38	1.21	1.08	0.97	0.97	0.90	0.78	0.71	0.62	0.48	0.41	0.34	0.25
	pow .8	7.84	4.83	3.63	2.96	2.57	2.26	2.04	1.86	1.74	1.62	1.43	1.25	1.06	0.82	0.69	0.55	0.39
	1% α=.05	13.49	7.44	5.43	4.42	3.78	3.35	3.07	2.85	2.68	2.54	2.32	2.11	1.90	1.68	1.56	1.44	1.30
	1% α=.01	19.02	10.26	7.33	5.87	4.98	4.39	3.97	3.65	3.40	3.20	2.90	2.60	2.29	1.97	1.80	1.63	1.44
	pow .5	13.21	6.73	4.56	3.47	2.79	2.46	2.14	1.89	1.70	1.54	1.38	1.12	0.92	0.68	0.56	0.43	0.30
	pow .8	20.18	10.35	7.07	5.42	4.40	3.77	3.29	2.93	2.65	2.42	2.10	1.75	1.41	1.04	0.86	0.66	0.44
	5% α=.05	39.07	20.15	13.84	10.68	8.79	7.53	6.62	5.95	5.42	4.98	4.33	3.71	3.08	2.44	2.12	1.80	1.48
	5% α=.01	48.68	24.94	17.05	13.11	10.74	9.16	8.04	7.19	6.53	6.00	5.21	4.42	3.63	2.84	2.44	2.04	1.63
	pow .5	38.66	19.38	12.95	9.73	7.80	6.51	5.59	4.90	4.36	4.04	3.37	2.71	2.04	1.42	1.11	0.80	0.48
	pow .8	49.91	25.07	16.78	12.64	10.16	8.50	7.32	6.43	5.73	5.24	4.39	3.55	2.71	1.88	1.47	1.06	0.63
500	nil α=.05	3.85	3.01	2.62	2.39	2.23	2.12	2.03	1.95	1.90	1.85	1.77	1.68	1.59	1.48	1.41	1.34	1.25
	nil α=.01	6.68	4.65	3.82	3.36	3.05	2.84	2.67	2.55	2.44	2.36	2.22	2.07	1.91	1.73	1.63	1.52	1.38
	pow .5	3.79	2.46	1.87	1.52	1.38	1.21	1.08	0.97	0.97	0.89	0.77	0.71	0.56	0.48	0.41	0.34	0.25
	pow .8	7.83	4.82	3.62	2.96	2.56	2.26	2.03	1.85	1.73	1.61	1.42	1.25	1.03	0.82	0.69	0.55	0.38
	1% α=.05	15.23	8.29	5.98	4.82	4.13	3.65	3.29	3.04	2.84	2.69	2.45	2.21	1.97	1.72	1.59	1.45	1.30
	1% α=.01	21.10	11.28	8.00	6.36	5.37	4.70	4.23	3.88	3.60	3.38	3.04	2.71	2.36	2.01	1.83	1.64	1.43
	pow .5	14.99	7.61	5.14	3.90	3.15	2.64	2.38	2.10	1.89	1.71	1.44	1.24	0.95	0.69	0.56	0.44	0.30
	pow .8	22.31	11.39	7.74	5.91	4.81	4.06	3.57	3.16	2.85	2.60	2.22	1.86	1.46	1.07	0.87	0.67	0.44
	5% α=.05	46.31	23.76	16.24	12.45	10.16	8.63	7.57	6.77	6.15	5.65	4.91	4.16	3.40	2.66	2.28	1.90	1.52
	5% α=.01	56.40	28.82	19.60	15.02	12.26	10.43	9.11	8.13	7.36	6.75	5.83	4.91	3.99	3.07	2.61	2.14	1.67
	pow .5	45.18	22.62	15.10	11.32	9.05	7.54	6.65	5.82	5.18	4.67	3.90	3.13	2.42	1.62	1.26	0.90	0.52
	pow .8	57.52	28.85	19.30	14.50	11.61	9.68	8.44	7.40	6.60	5.96	4.99	4.03	3.09	2.11	1.64	1.17	0.67
600	nil α=.05	3.85	3.01	2.62	2.39	2.23	2.11	2.02	1.95	1.89	1.84	1.77	1.68	1.58	1.47	1.41	1.34	1.25
	nil α=.01	6.67	4.64	3.81	3.35	3.05	2.83	2.67	2.54	2.44	2.35	2.21	2.07	1.91	1.73	1.63	1.51	1.37
	pow .5	3.79	2.46	1.87	1.52	1.38	1.21	1.08	0.97	0.96	0.89	0.77	0.71	0.56	0.48	0.41	0.34	0.25
	pow .8	7.82	4.82	3.62	2.95	2.56	2.25	2.02	1.85	1.73	1.61	1.42	1.24	1.02	0.81	0.68	0.54	0.38
	1% α=.05	16.94	9.11	6.51	5.21	4.43	3.91	3.54	3.24	3.01	2.84	2.57	2.31	2.03	1.76	1.62	1.47	1.31
	1% α=.01	23.08	12.25	8.64	6.83	5.74	5.01	4.49	4.10	3.80	3.55	3.18	2.81	2.44	2.06	1.86	1.66	1.44
	pow .5	16.14	8.46	5.71	4.32	3.49	2.93	2.52	2.32	2.07	1.88	1.59	1.28	1.04	0.75	0.61	0.47	0.30
	pow .8	24.06	12.38	8.38	6.38	5.18	4.38	3.80	3.41	3.05	2.78	2.37	1.95	1.55	1.13	0.91	0.69	0.44
	5% α=.05	52.82	26.98	18.38	14.08	11.50	9.78	8.55	7.63	6.91	6.34	5.48	4.59	3.73	2.86	2.44	2.00	1.56
	5% α=.01	63.87	32.55	22.11	16.89	13.75	11.65	10.16	9.05	8.18	7.48	6.44	5.39	4.35	3.30	2.78	2.25	1.72
	pow .5	52.51	26.28	17.54	13.17	10.55	8.80	7.55	6.61	5.88	5.30	4.42	3.63	2.73	1.88	1.41	1.00	0.57
	pow .8	65.29	32.72	21.87	16.44	13.19	11.02	9.47	8.30	7.40	6.67	5.59	4.54	3.44	2.37	1.80	1.27	0.72

One Stop F Table

	Hyp F									df_{hyp}								
df_{err}	for	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
1000	nil α=.05	3.84	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.76	1.67	1.58	1.47	1.40	1.33	1.24
	nil α=.01	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.42	2.34	2.20	2.06	1.90	1.71	1.61	1.49	1.35
	pow .5	3.79	2.46	1.87	1.51	1.37	1.20	1.07	0.97	0.96	0.89	0.77	0.71	0.55	0.48	0.41	0.30	0.23
	pow .8	7.81	4.81	3.61	2.94	2.55	2.24	2.02	1.84	1.72	1.60	1.41	1.23	1.01	0.80	0.67	0.52	0.36
	1% α=.05	23.25	12.26	8.59	6.76	5.66	4.93	4.40	3.99	3.66	3.42	3.05	2.68	2.30	1.93	1.74	1.54	1.34
	1% α=.01	30.44	15.89	11.01	8.59	7.13	6.16	5.47	4.95	4.54	4.22	3.73	3.24	2.75	2.25	2.00	1.74	1.46
	pow .5	22.91	11.53	7.73	5.83	4.68	3.92	3.37	3.08	2.73	2.47	2.07	1.67	1.33	0.95	0.72	0.54	0.33
	pow .8	31.72	16.01	10.78	8.16	6.58	5.53	4.78	4.27	3.81	3.45	2.91	2.38	1.86	1.33	1.04	0.76	0.47
	5% α=.05	78.99	40.07	27.09	20.61	16.71	14.12	12.27	10.88	9.79	8.93	7.63	6.32	5.00	3.71	3.06	2.41	1.75
	5% α=.01	92.43	46.81	31.60	23.96	19.41	16.37	14.20	12.57	11.30	10.29	8.77	7.25	5.73	4.21	3.45	2.68	1.92
	pow .5	78.54	39.29	26.20	19.66	15.74	13.12	11.25	9.85	8.75	7.88	6.57	5.36	4.02	2.69	2.06	1.40	0.75
	pow .8	93.82	46.97	31.35	23.54	18.86	15.73	13.50	11.83	10.53	9.48	7.92	6.42	4.83	3.25	2.49	1.70	0.92
10000	nil α=.05	3.84	3.00	2.61	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.66	1.57	1.46	1.39	1.32	1.22
	nil α=.01	6.64	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.19	2.04	1.88	1.70	1.59	1.47	1.33
	pow .5	3.79	2.34	1.86	1.51	1.37	1.20	1.07	0.96	0.87	0.84	0.77	0.63	0.55	0.43	0.36	0.30	0.22
	pow .8	7.81	4.76	3.60	2.93	2.54	2.23	2.00	1.83	1.68	1.59	1.40	1.20	1.00	0.77	0.64	0.51	0.34
	1% α=.05	135.8	68.43	45.99	34.77	28.04	23.55	20.34	17.94	16.07	14.57	12.33	10.06	7.80	5.56	4.44	3.31	2.19
	1% α=.01	152.7	76.89	51.63	38.99	31.39	26.35	22.74	20.04	17.94	16.26	13.73	11.21	8.69	6.16	4.90	3.63	2.36
	pow .5	134.7	67.36	44.90	33.68	26.95	22.46	19.25	16.85	14.98	13.48	11.23	8.99	6.74	4.55	3.41	2.31	1.19
	pow .8	154.1	77.06	51.39	38.56	30.86	25.73	22.06	19.31	17.17	15.46	12.90	10.32	7.75	5.22	3.93	2.66	1.37
	5% α=.05	601.3	301.2	201.2	151.1	121.1	101.1	86.81	76.09	67.76	61.09	51.08	41.08	31.07	21.06	16.05	11.04	6.03
	5% α=.01	637.8	319.4	213.3	160.3	128.4	107.2	92.03	80.66	71.81	64.74	54.12	43.51	32.89	22.28	16.97	11.67	6.36
	pow .5	600.6	300.3	200.1	150.1	120.1	100.1	85.79	75.05	66.72	60.06	50.04	40.02	30.01	20.01	15.01	10.04	5.03
	pow .8	639.9	319.5	213.1	159.9	127.7	106.6	91.23	79.99	71.07	64.00	53.31	42.68	31.94	21.33	16.00	10.68	5.38

Appendix C

One-Stop *PV* Table

One-Stop PV Table

										df_{hyp}						
df_{err}		1	2	3	4	5	6	7	8	9	10	12	15	20	30	40
3	nil .05	0.771	0.864	0.903	0.924	0.938	0.947	0.954	0.959	0.964	0.967	0.972	0.978	0.983	0.989	0.991
	nil .01	0.919	0.954	0.967	0.975	0.979	0.982	0.985	0.987	0.988	0.989	0.991	0.993	0.994	0.996	0.997
	pow .5	0.734	0.828	0.871	0.897	0.914	0.926	0.936	0.943	0.948	0.953	0.960	0.968	0.975	0.983	0.987
	pow .8	0.858	0.913	0.937	0.951	0.959	0.965	0.970	0.973	0.976	0.978	0.982	0.985	0.989	0.993	0.994
	1% .05	0.777	0.866	0.904	0.925	0.938	0.947	0.954	0.959	0.964	0.967	0.972	0.978	0.983	0.989	0.991
	1% .01	0.921	0.954	0.967	0.975	0.979	0.982	0.985	0.987	0.988	0.989	0.991	0.993	0.994	0.996	0.997
	pow .5	0.740	0.830	0.873	0.898	0.915	0.927	0.936	0.943	0.948	0.953	0.960	0.968	0.975	0.983	0.987
	pow .8	0.862	0.914	0.937	0.951	0.960	0.965	0.970	0.974	0.976	0.979	0.982	0.985	0.989	0.993	0.994
	5% .05	0.796	0.873	0.907	0.927	0.939	0.948	0.955	0.960	0.964	0.967	0.973	0.978	0.983	0.989	0.991
	5% .01	0.929	0.957	0.969	0.975	0.980	0.983	0.985	0.987	0.988	0.989	0.991	0.993	0.994	0.996	0.997
	pow .5	0.761	0.839	0.878	0.901	0.917	0.928	0.937	0.944	0.949	0.954	0.961	0.968	0.975	0.983	0.987
	pow .8	0.874	0.919	0.940	0.952	0.961	0.966	0.971	0.974	0.977	0.979	0.982	0.986	0.989	0.993	0.994
4	nil .05	0.658	0.776	0.832	0.865	0.887	0.902	0.914	0.924	0.931	0.937	0.947	0.956	0.967	0.977	0.983
	nil .01	0.841	0.900	0.926	0.941	0.951	0.958	0.963	0.967	0.971	0.973	0.977	0.982	0.986	0.990	0.993
	pow .5	0.625	0.733	0.789	0.825	0.850	0.869	0.884	0.895	0.905	0.912	0.925	0.938	0.952	0.967	0.975
	pow .8	0.780	0.850	0.885	0.906	0.920	0.931	0.939	0.946	0.951	0.955	0.962	0.969	0.976	0.984	0.988
	1% .05	0.667	0.780	0.834	0.866	0.887	0.903	0.915	0.924	0.931	0.937	0.947	0.957	0.967	0.977	0.983
	1% .01	0.846	0.902	0.927	0.942	0.951	0.958	0.963	0.967	0.971	0.973	0.977	0.982	0.986	0.990	0.993
	pow .5	0.634	0.737	0.792	0.827	0.852	0.870	0.884	0.896	0.905	0.913	0.925	0.938	0.952	0.967	0.975
	pow .8	0.785	0.852	0.886	0.907	0.921	0.932	0.940	0.946	0.951	0.955	0.962	0.969	0.976	0.984	0.988
	5% .05	0.699	0.793	0.841	0.871	0.891	0.905	0.917	0.925	0.932	0.938	0.947	0.957	0.967	0.977	0.983
	5% .01	0.864	0.909	0.931	0.944	0.953	0.959	0.964	0.968	0.971	0.974	0.978	0.982	0.986	0.991	0.993
	pow .5	0.668	0.753	0.802	0.834	0.856	0.874	0.887	0.898	0.907	0.914	0.926	0.939	0.953	0.967	0.975
	pow .8	0.806	0.861	0.891	0.911	0.923	0.933	0.941	0.947	0.952	0.956	0.963	0.969	0.976	0.984	0.988

One-Stop PV Table

										df_{hyp}						
df_{err}		1	2	3	4	5	6	7	8	9	10	12	15	20	30	40
5	nil .05	0.569	0.698	0.764	0.806	0.835	0.856	0.872	0.885	0.896	0.904	0.918	0.933	0.948	0.964	0.973
	nil .01	0.765	0.842	0.879	0.901	0.916	0.928	0.936	0.943	0.948	0.953	0.960	0.967	0.974	0.983	0.987
	pow .5	0.542	0.651	0.713	0.756	0.787	0.811	0.830	0.845	0.858	0.869	0.886	0.904	0.925	0.947	0.959
	pow .8	0.712	0.790	0.831	0.859	0.879	0.893	0.905	0.913	0.921	0.928	0.938	0.948	0.960	0.972	0.979
	1% .05	0.581	0.703	0.767	0.808	0.836	0.857	0.873	0.886	0.896	0.905	0.919	0.933	0.948	0.964	0.973
	1% .01	0.773	0.845	0.880	0.902	0.917	0.928	0.936	0.943	0.948	0.953	0.960	0.967	0.975	0.983	0.987
	pow .5	0.553	0.657	0.718	0.759	0.789	0.812	0.831	0.846	0.859	0.869	0.886	0.905	0.925	0.947	0.959
	pow .8	0.720	0.793	0.834	0.860	0.880	0.894	0.905	0.914	0.922	0.928	0.938	0.949	0.960	0.972	0.979
	5% .05	0.624	0.723	0.779	0.816	0.842	0.861	0.876	0.888	0.898	0.907	0.920	0.934	0.949	0.965	0.973
	5% .01	0.802	0.857	0.887	0.907	0.920	0.930	0.938	0.944	0.950	0.954	0.960	0.967	0.975	0.983	0.987
	pow .5	0.597	0.681	0.732	0.770	0.798	0.819	0.836	0.850	0.862	0.872	0.889	0.906	0.926	0.948	0.960
	pow .8	0.749	0.808	0.842	0.866	0.884	0.897	0.908	0.916	0.924	0.930	0.939	0.949	0.960	0.972	0.979
6	nil .05	0.499	0.632	0.704	0.751	0.785	0.811	0.831	0.847	0.860	0.871	0.889	0.908	0.928	0.950	0.962
	nil .01	0.696	0.785	0.830	0.859	0.879	0.894	0.906	0.915	0.923	0.929	0.939	0.950	0.961	0.973	0.979
	pow .5	0.476	0.581	0.647	0.693	0.728	0.755	0.777	0.796	0.811	0.824	0.845	0.869	0.896	0.926	0.942
	pow .8	0.654	0.734	0.781	0.813	0.836	0.854	0.868	0.880	0.890	0.898	0.911	0.925	0.941	0.959	0.968
	1% .05	0.514	0.638	0.708	0.754	0.787	0.812	0.832	0.848	0.861	0.872	0.889	0.908	0.928	0.950	0.962
	1% .01	0.708	0.790	0.833	0.861	0.881	0.895	0.907	0.916	0.923	0.930	0.939	0.950	0.961	0.973	0.979
	pow .5	0.490	0.590	0.652	0.697	0.731	0.757	0.779	0.797	0.812	0.825	0.847	0.870	0.896	0.926	0.942
	pow .8	0.664	0.740	0.784	0.815	0.838	0.855	0.869	0.881	0.890	0.898	0.912	0.926	0.941	0.959	0.968
	5% .05	0.566	0.664	0.724	0.765	0.795	0.818	0.837	0.852	0.864	0.875	0.891	0.910	0.929	0.951	0.962
	5% .01	0.747	0.807	0.844	0.868	0.886	0.899	0.910	0.918	0.925	0.931	0.941	0.951	0.962	0.973	0.980
	pow .5	0.542	0.619	0.674	0.713	0.743	0.767	0.787	0.803	0.817	0.830	0.850	0.872	0.898	0.927	0.943
	pow .8	0.701	0.759	0.798	0.824	0.845	0.861	0.874	0.884	0.893	0.901	0.913	0.927	0.942	0.959	0.968

One-Stop PV Table

		df_{hyp}														
df_{err}		1	2	3	4	5	6	7	8	9	10	12	15	20	30	40
8	nil .05	0.399	0.527	0.604	0.657	0.697	0.729	0.754	0.775	0.792	0.807	0.831	0.858	0.887	0.920	0.938
	nil .01	0.585	0.684	0.740	0.778	0.806	0.827	0.844	0.858	0.869	0.879	0.895	0.912	0.931	0.951	0.962
	pow .5	0.382	0.476	0.539	0.585	0.624	0.654	0.679	0.703	0.721	0.737	0.766	0.798	0.836	0.880	0.905
	pow .8	0.561	0.642	0.692	0.727	0.756	0.777	0.795	0.811	0.824	0.835	0.854	0.875	0.900	0.928	0.943
	1% .05	0.418	0.537	0.610	0.662	0.701	0.731	0.756	0.776	0.794	0.808	0.832	0.858	0.888	0.920	0.938
	1% .01	0.603	0.692	0.745	0.781	0.808	0.829	0.845	0.859	0.870	0.880	0.895	0.912	0.931	0.951	0.962
	pow .5	0.401	0.487	0.547	0.591	0.629	0.657	0.682	0.705	0.723	0.741	0.767	0.799	0.837	0.880	0.905
	pow .8	0.576	0.650	0.697	0.731	0.759	0.780	0.797	0.813	0.825	0.837	0.855	0.876	0.900	0.928	0.943
	5% .05	0.482	0.573	0.634	0.679	0.714	0.742	0.764	0.783	0.800	0.813	0.836	0.861	0.889	0.921	0.939
	5% .01	0.658	0.721	0.764	0.794	0.818	0.836	0.851	0.864	0.874	0.883	0.898	0.914	0.932	0.952	0.963
	pow .5	0.465	0.528	0.578	0.615	0.646	0.674	0.696	0.717	0.734	0.749	0.775	0.805	0.840	0.882	0.906
	pow .8	0.625	0.679	0.718	0.747	0.770	0.790	0.806	0.820	0.832	0.842	0.859	0.879	0.902	0.929	0.944
10	nil .05	0.332	0.451	0.527	0.582	0.624	0.659	0.687	0.711	0.731	0.749	0.778	0.810	0.847	0.890	0.914
	nil .01	0.501	0.602	0.663	0.706	0.738	0.764	0.784	0.802	0.816	0.829	0.850	0.872	0.898	0.927	0.943
	pow .5	0.319	0.400	0.459	0.503	0.539	0.570	0.596	0.620	0.640	0.659	0.691	0.731	0.776	0.833	0.866
	pow .8	0.491	0.568	0.618	0.655	0.684	0.708	0.729	0.746	0.761	0.775	0.798	0.825	0.856	0.893	0.915
	1% .05	0.353	0.463	0.535	0.588	0.629	0.662	0.690	0.713	0.733	0.750	0.779	0.811	0.848	0.890	0.914
	1% .01	0.524	0.613	0.670	0.711	0.742	0.767	0.787	0.804	0.818	0.830	0.851	0.873	0.899	0.927	0.944
	pow .5	0.339	0.416	0.468	0.510	0.545	0.575	0.600	0.623	0.643	0.665	0.696	0.733	0.778	0.833	0.867
	pow .8	0.509	0.579	0.626	0.661	0.689	0.712	0.732	0.749	0.764	0.778	0.800	0.826	0.856	0.893	0.915
	5% .05	0.425	0.507	0.566	0.611	0.647	0.677	0.702	0.724	0.742	0.758	0.785	0.815	0.851	0.892	0.915
	5% .01	0.591	0.652	0.696	0.730	0.756	0.778	0.796	0.812	0.825	0.836	0.855	0.876	0.900	0.928	0.944
	pow .5	0.410	0.465	0.507	0.542	0.572	0.598	0.621	0.641	0.659	0.676	0.705	0.739	0.784	0.837	0.869
	pow .8	0.568	0.617	0.653	0.682	0.706	0.727	0.744	0.760	0.773	0.785	0.806	0.830	0.860	0.895	0.916

One-Stop PV Table

									df_{hyp}							
df_{err}		1	2	3	4	5	6	7	8	9	10	12	15	20	30	40
12	nil .05	0.283	0.393	0.466	0.521	0.564	0.600	0.630	0.655	0.677	0.696	0.729	0.766	0.809	0.860	0.890
	nil .01	0.437	0.536	0.598	0.643	0.678	0.707	0.730	0.750	0.767	0.782	0.806	0.834	0.865	0.902	0.923
	pow .5	0.274	0.346	0.396	0.437	0.472	0.502	0.529	0.553	0.575	0.595	0.625	0.667	0.718	0.786	0.826
	pow .8	0.436	0.509	0.557	0.594	0.623	0.648	0.670	0.689	0.706	0.721	0.744	0.775	0.811	0.858	0.885
	1% .05	0.307	0.407	0.476	0.528	0.570	0.604	0.634	0.658	0.680	0.699	0.731	0.767	0.810	0.861	0.890
	1% .01	0.464	0.550	0.607	0.650	0.684	0.711	0.734	0.753	0.769	0.784	0.808	0.835	0.866	0.903	0.924
	pow .5	0.296	0.360	0.408	0.446	0.479	0.508	0.534	0.558	0.579	0.598	0.627	0.670	0.719	0.787	0.827
	pow .8	0.457	0.522	0.566	0.601	0.629	0.653	0.674	0.692	0.709	0.723	0.747	0.777	0.812	0.858	0.885
	5% .05	0.384	0.457	0.512	0.556	0.592	0.623	0.649	0.672	0.692	0.709	0.739	0.773	0.814	0.863	0.891
	5% .01	0.540	0.597	0.640	0.675	0.703	0.727	0.747	0.764	0.779	0.792	0.814	0.839	0.869	0.904	0.925
	pow .5	0.371	0.415	0.452	0.483	0.512	0.537	0.560	0.580	0.599	0.617	0.643	0.682	0.728	0.792	0.830
	pow .8	0.523	0.566	0.600	0.628	0.652	0.673	0.691	0.707	0.722	0.735	0.756	0.784	0.818	0.861	0.887
14	nil .05	0.247	0.348	0.417	0.471	0.514	0.550	0.580	0.607	0.630	0.650	0.685	0.725	0.773	0.832	0.866
	nil .01	0.388	0.482	0.544	0.590	0.626	0.656	0.681	0.703	0.721	0.738	0.765	0.797	0.834	0.878	0.903
	pow .5	0.240	0.304	0.351	0.389	0.416	0.446	0.472	0.496	0.518	0.532	0.569	0.610	0.666	0.738	0.785
	pow .8	0.393	0.461	0.507	0.543	0.571	0.596	0.618	0.637	0.655	0.669	0.696	0.728	0.769	0.821	0.854
	1% .05	0.272	0.364	0.429	0.479	0.521	0.555	0.585	0.611	0.633	0.653	0.687	0.727	0.774	0.832	0.867
	1% .01	0.418	0.499	0.555	0.598	0.633	0.662	0.686	0.706	0.725	0.740	0.767	0.798	0.835	0.878	0.904
	pow .5	0.264	0.320	0.363	0.399	0.430	0.453	0.478	0.501	0.523	0.543	0.572	0.612	0.668	0.739	0.786
	pow .8	0.416	0.476	0.518	0.551	0.579	0.602	0.623	0.642	0.659	0.674	0.699	0.730	0.770	0.822	0.854
	5% .05	0.353	0.418	0.469	0.511	0.547	0.577	0.604	0.627	0.648	0.666	0.697	0.735	0.780	0.835	0.868
	5% .01	0.499	0.552	0.594	0.628	0.657	0.681	0.702	0.721	0.737	0.751	0.776	0.804	0.838	0.880	0.905
	pow .5	0.341	0.379	0.411	0.440	0.467	0.485	0.508	0.528	0.547	0.559	0.592	0.628	0.679	0.746	0.791
	pow .8	0.488	0.526	0.557	0.583	0.607	0.625	0.644	0.660	0.675	0.687	0.711	0.740	0.778	0.826	0.857

One-Stop PV Table

									df_{hyp}							
df_{err}		1	2	3	4	5	6	7	8	9	10	12	15	20	30	40
16	nil .05	0.219	0.312	0.378	0.429	0.471	0.507	0.538	0.564	0.588	0.609	0.645	0.688	0.740	0.804	0.843
	nil .01	0.348	0.438	0.498	0.544	0.581	0.612	0.638	0.660	0.680	0.698	0.727	0.762	0.803	0.853	0.883
	pow .5	0.213	0.271	0.310	0.345	0.376	0.403	0.422	0.446	0.468	0.489	0.519	0.561	0.615	0.696	0.746
	pow .8	0.357	0.421	0.464	0.498	0.527	0.551	0.571	0.591	0.609	0.625	0.652	0.685	0.728	0.786	0.823
	1% .05	0.245	0.329	0.390	0.439	0.479	0.513	0.543	0.569	0.592	0.613	0.648	0.690	0.741	0.805	0.844
	1% .01	0.380	0.456	0.511	0.554	0.589	0.618	0.643	0.665	0.684	0.701	0.730	0.764	0.804	0.854	0.883
	pow .5	0.238	0.289	0.328	0.361	0.385	0.411	0.436	0.452	0.473	0.493	0.523	0.564	0.617	0.697	0.747
	pow .8	0.382	0.437	0.477	0.509	0.535	0.558	0.579	0.596	0.613	0.629	0.655	0.688	0.730	0.788	0.824
	5% .05	0.328	0.387	0.434	0.475	0.509	0.539	0.565	0.588	0.609	0.628	0.660	0.699	0.748	0.809	0.846
	5% .01	0.465	0.514	0.554	0.588	0.617	0.641	0.663	0.682	0.699	0.714	0.740	0.771	0.809	0.857	0.885
	pow .5	0.318	0.350	0.379	0.406	0.425	0.447	0.469	0.482	0.501	0.519	0.546	0.583	0.632	0.706	0.753
	pow .8	0.458	0.492	0.521	0.546	0.566	0.586	0.604	0.618	0.633	0.647	0.670	0.700	0.739	0.793	0.827
18	nil .05	0.197	0.283	0.345	0.394	0.435	0.470	0.500	0.527	0.551	0.573	0.610	0.654	0.709	0.778	0.821
	nil .01	0.315	0.401	0.459	0.504	0.541	0.572	0.599	0.622	0.643	0.661	0.692	0.729	0.774	0.829	0.863
	pow .5	0.190	0.242	0.280	0.313	0.342	0.362	0.386	0.409	0.423	0.444	0.474	0.517	0.573	0.654	0.706
	pow .8	0.327	0.387	0.428	0.461	0.488	0.511	0.532	0.552	0.567	0.584	0.611	0.646	0.691	0.752	0.792
	1% .05	0.224	0.301	0.358	0.405	0.444	0.477	0.507	0.533	0.556	0.577	0.613	0.657	0.711	0.779	0.822
	1% .01	0.349	0.421	0.473	0.515	0.550	0.580	0.605	0.627	0.647	0.665	0.695	0.731	0.775	0.830	0.864
	pow .5	0.218	0.263	0.299	0.324	0.352	0.377	0.394	0.416	0.437	0.449	0.478	0.520	0.575	0.656	0.711
	pow .8	0.354	0.405	0.443	0.471	0.497	0.521	0.539	0.558	0.575	0.589	0.615	0.649	0.693	0.754	0.794
	5% .05	0.308	0.361	0.406	0.444	0.477	0.506	0.531	0.554	0.575	0.594	0.627	0.668	0.718	0.784	0.824
	5% .01	0.437	0.483	0.521	0.553	0.582	0.606	0.628	0.647	0.665	0.680	0.708	0.741	0.782	0.834	0.866
	pow .5	0.298	0.327	0.353	0.372	0.394	0.416	0.430	0.449	0.467	0.478	0.504	0.542	0.592	0.667	0.715
	pow .8	0.433	0.463	0.490	0.511	0.532	0.551	0.567	0.583	0.598	0.610	0.633	0.664	0.704	0.761	0.798

One-Stop PV Table

										df_{hyp}						
df_{err}		1	2	3	4	5	6	7	8	9	10	12	15	20	30	40
20	nil .05	0.178	0.259	0.317	0.364	0.404	0.438	0.468	0.495	0.518	0.540	0.577	0.623	0.680	0.754	0.799
	nil .01	0.288	0.369	0.426	0.470	0.506	0.537	0.564	0.588	0.609	0.627	0.660	0.698	0.746	0.806	0.843
	pow .5	0.172	0.220	0.256	0.286	0.307	0.332	0.356	0.370	0.391	0.403	0.441	0.475	0.533	0.612	0.669
	pow .8	0.301	0.358	0.397	0.428	0.453	0.477	0.498	0.515	0.532	0.546	0.576	0.609	0.655	0.719	0.762
	1% .05	0.206	0.277	0.331	0.376	0.413	0.446	0.475	0.501	0.524	0.545	0.581	0.626	0.682	0.755	0.800
	1% .01	0.324	0.391	0.441	0.482	0.516	0.545	0.571	0.594	0.614	0.632	0.663	0.701	0.748	0.808	0.844
	pow .5	0.201	0.242	0.270	0.298	0.324	0.341	0.364	0.385	0.398	0.418	0.446	0.479	0.536	0.614	0.675
	pow .8	0.330	0.378	0.411	0.440	0.465	0.485	0.505	0.524	0.538	0.554	0.580	0.613	0.658	0.720	0.764
	5% .05	0.291	0.340	0.381	0.417	0.449	0.477	0.502	0.525	0.545	0.564	0.597	0.639	0.691	0.760	0.804
	5% .01	0.414	0.456	0.492	0.524	0.551	0.575	0.597	0.616	0.634	0.650	0.678	0.712	0.756	0.812	0.847
	pow .5	0.285	0.307	0.331	0.348	0.369	0.389	0.402	0.420	0.431	0.448	0.473	0.503	0.555	0.627	0.679
	pow .8	0.412	0.439	0.463	0.483	0.503	0.521	0.536	0.552	0.564	0.578	0.601	0.630	0.671	0.729	0.769
22	nil .05	0.163	0.238	0.294	0.339	0.377	0.410	0.439	0.466	0.489	0.511	0.548	0.594	0.653	0.730	0.779
	nil .01	0.265	0.342	0.396	0.440	0.475	0.506	0.533	0.557	0.578	0.597	0.630	0.670	0.720	0.784	0.824
	pow .5	0.158	0.202	0.235	0.263	0.283	0.307	0.322	0.343	0.356	0.375	0.404	0.446	0.495	0.578	0.638
	pow .8	0.280	0.333	0.370	0.400	0.424	0.447	0.465	0.484	0.499	0.515	0.541	0.577	0.622	0.688	0.734
	1% .05	0.192	0.258	0.308	0.351	0.387	0.419	0.447	0.472	0.495	0.516	0.553	0.598	0.655	0.731	0.780
	1% .01	0.302	0.365	0.413	0.453	0.486	0.515	0.541	0.563	0.584	0.602	0.634	0.673	0.722	0.786	0.825
	pow .5	0.187	0.224	0.250	0.276	0.301	0.317	0.338	0.351	0.371	0.382	0.409	0.451	0.499	0.579	0.639
	pow .8	0.310	0.354	0.386	0.413	0.437	0.456	0.476	0.491	0.508	0.521	0.547	0.581	0.625	0.690	0.735
	5% .05	0.277	0.322	0.360	0.394	0.425	0.452	0.476	0.498	0.519	0.537	0.571	0.612	0.665	0.737	0.784
	5% .01	0.393	0.433	0.467	0.497	0.524	0.548	0.569	0.588	0.606	0.622	0.651	0.686	0.731	0.791	0.829
	pow .5	0.271	0.291	0.312	0.327	0.347	0.359	0.378	0.388	0.406	0.415	0.439	0.476	0.519	0.594	0.649
	pow .8	0.393	0.418	0.440	0.459	0.478	0.493	0.509	0.522	0.536	0.547	0.570	0.600	0.640	0.700	0.742

One-Stop PV Table

		df_{hyp}														
df_{err}		1	2	3	4	5	6	7	8	9	10	12	15	20	30	40
24	nil .05	0.151	0.221	0.273	0.316	0.353	0.385	0.414	0.440	0.463	0.484	0.522	0.568	0.628	0.708	0.759
	nil .01	0.246	0.319	0.371	0.413	0.448	0.478	0.505	0.528	0.550	0.569	0.603	0.644	0.695	0.763	0.806
	pow .5	0.146	0.187	0.218	0.244	0.263	0.285	0.299	0.320	0.332	0.351	0.378	0.411	0.460	0.544	0.600
	pow .8	0.261	0.312	0.347	0.376	0.398	0.420	0.438	0.456	0.471	0.486	0.513	0.545	0.591	0.659	0.705
	1% .05	0.180	0.241	0.289	0.329	0.364	0.395	0.422	0.447	0.470	0.490	0.527	0.572	0.631	0.709	0.760
	1% .01	0.284	0.343	0.389	0.427	0.459	0.488	0.513	0.536	0.556	0.575	0.607	0.647	0.698	0.765	0.807
	pow .5	0.175	0.210	0.233	0.258	0.274	0.296	0.308	0.328	0.347	0.358	0.384	0.416	0.473	0.546	0.609
	pow .8	0.292	0.333	0.363	0.389	0.410	0.430	0.447	0.464	0.481	0.493	0.519	0.550	0.597	0.661	0.709
	5% .05	0.264	0.306	0.342	0.375	0.403	0.430	0.453	0.475	0.495	0.513	0.546	0.588	0.642	0.716	0.765
	5% .01	0.376	0.413	0.445	0.474	0.500	0.523	0.544	0.563	0.581	0.597	0.626	0.662	0.708	0.770	0.811
	pow .5	0.259	0.276	0.296	0.310	0.328	0.340	0.357	0.367	0.384	0.392	0.415	0.452	0.495	0.562	0.621
	pow .8	0.377	0.399	0.420	0.437	0.455	0.469	0.485	0.497	0.511	0.522	0.543	0.574	0.613	0.672	0.717
26	nil .05	0.140	0.206	0.256	0.297	0.332	0.363	0.391	0.417	0.439	0.460	0.498	0.544	0.605	0.687	0.740
	nil .01	0.229	0.298	0.349	0.389	0.423	0.453	0.479	0.503	0.524	0.543	0.577	0.619	0.672	0.743	0.788
	pow .5	0.135	0.174	0.202	0.222	0.245	0.259	0.280	0.291	0.311	0.320	0.347	0.388	0.436	0.512	0.570
	pow .8	0.245	0.293	0.326	0.352	0.376	0.395	0.414	0.429	0.445	0.458	0.484	0.519	0.564	0.631	0.679
	1% .05	0.169	0.226	0.271	0.310	0.343	0.373	0.400	0.424	0.446	0.467	0.503	0.549	0.608	0.689	0.741
	1% .01	0.268	0.323	0.367	0.404	0.436	0.464	0.488	0.511	0.531	0.550	0.582	0.623	0.675	0.744	0.789
	pow .5	0.165	0.197	0.218	0.241	0.257	0.277	0.289	0.308	0.318	0.336	0.362	0.393	0.440	0.515	0.580
	pow .8	0.277	0.315	0.343	0.368	0.388	0.407	0.423	0.440	0.453	0.468	0.493	0.524	0.568	0.633	0.683
	5% .05	0.254	0.293	0.327	0.357	0.385	0.410	0.433	0.454	0.474	0.492	0.524	0.566	0.620	0.696	0.747
	5% .01	0.360	0.395	0.426	0.453	0.478	0.501	0.522	0.540	0.558	0.574	0.603	0.639	0.686	0.751	0.794
	pow .5	0.248	0.264	0.282	0.295	0.312	0.322	0.339	0.348	0.364	0.372	0.395	0.421	0.463	0.540	0.593
	pow .8	0.362	0.383	0.402	0.418	0.435	0.448	0.463	0.475	0.488	0.498	0.520	0.547	0.586	0.648	0.692

One-Stop PV Table

									df_{hyp}							
df_{err}		1	2	3	4	5	6	7	8	9	10	12	15	20	30	40
28	nil .05	0.130	0.193	0.240	0.279	0.314	0.344	0.371	0.396	0.418	0.439	0.476	0.522	0.583	0.667	0.722
	nil .01	0.214	0.280	0.329	0.368	0.401	0.430	0.456	0.480	0.501	0.520	0.554	0.596	0.650	0.723	0.771
	pow .5	0.126	0.162	0.189	0.208	0.229	0.243	0.263	0.273	0.292	0.301	0.327	0.357	0.405	0.491	0.541
	pow .8	0.230	0.276	0.308	0.333	0.355	0.373	0.392	0.407	0.422	0.435	0.460	0.492	0.536	0.607	0.653
	1% .05	0.160	0.213	0.256	0.293	0.325	0.354	0.380	0.404	0.426	0.446	0.481	0.527	0.586	0.669	0.724
	1% .01	0.254	0.306	0.348	0.383	0.414	0.442	0.466	0.488	0.508	0.527	0.559	0.600	0.653	0.725	0.772
	pow .5	0.156	0.185	0.206	0.227	0.241	0.261	0.272	0.291	0.300	0.318	0.333	0.373	0.419	0.493	0.551
	pow .8	0.263	0.299	0.325	0.349	0.368	0.387	0.402	0.418	0.431	0.446	0.467	0.500	0.544	0.609	0.658
	5% .05	0.245	0.281	0.313	0.342	0.368	0.392	0.415	0.435	0.454	0.472	0.504	0.545	0.600	0.677	0.729
	5% .01	0.346	0.379	0.409	0.435	0.459	0.481	0.501	0.520	0.537	0.553	0.581	0.618	0.666	0.733	0.777
	pow .5	0.239	0.253	0.270	0.281	0.298	0.307	0.323	0.331	0.339	0.354	0.376	0.402	0.443	0.511	0.565
	pow .8	0.349	0.368	0.386	0.401	0.417	0.430	0.444	0.455	0.465	0.478	0.498	0.525	0.563	0.623	0.668
30	nil .05	0.122	0.181	0.226	0.264	0.297	0.326	0.353	0.377	0.399	0.419	0.455	0.502	0.563	0.648	0.705
	nil .01	0.201	0.264	0.311	0.349	0.381	0.410	0.435	0.458	0.479	0.498	0.532	0.574	0.629	0.705	0.754
	pow .5	0.118	0.152	0.178	0.195	0.216	0.228	0.247	0.258	0.276	0.284	0.309	0.339	0.385	0.461	0.522
	pow .8	0.217	0.261	0.292	0.315	0.337	0.354	0.372	0.386	0.402	0.414	0.438	0.469	0.513	0.581	0.631
	1% .05	0.152	0.202	0.243	0.278	0.309	0.337	0.362	0.385	0.407	0.426	0.462	0.507	0.566	0.650	0.706
	1% .01	0.241	0.291	0.331	0.365	0.395	0.422	0.446	0.467	0.487	0.506	0.538	0.579	0.633	0.707	0.755
	pow .5	0.149	0.176	0.194	0.215	0.228	0.247	0.257	0.275	0.284	0.292	0.316	0.344	0.389	0.464	0.523
	pow .8	0.251	0.285	0.310	0.332	0.350	0.368	0.383	0.399	0.411	0.422	0.446	0.476	0.518	0.584	0.633
	5% .05	0.236	0.271	0.300	0.328	0.353	0.377	0.398	0.418	0.437	0.454	0.485	0.526	0.581	0.659	0.713
	5% .01	0.334	0.365	0.393	0.418	0.442	0.463	0.483	0.501	0.518	0.533	0.562	0.598	0.647	0.715	0.761
	pow .5	0.231	0.243	0.259	0.269	0.285	0.294	0.309	0.317	0.324	0.339	0.359	0.385	0.425	0.483	0.539
	pow .8	0.338	0.355	0.372	0.386	0.401	0.413	0.426	0.437	0.447	0.459	0.479	0.505	0.542	0.599	0.645

One-Stop PV Table

		df_{hyp}														
df_{err}		1	2	3	4	5	6	7	8	9	10	12	15	20	30	40
40	nil .05	0.092	0.139	0.176	0.207	0.234	0.259	0.282	0.304	0.323	0.342	0.375	0.419	0.479	0.567	0.628
	nil .01	0.155	0.206	0.244	0.277	0.305	0.331	0.353	0.374	0.394	0.412	0.444	0.486	0.542	0.623	0.679
	pow .5	0.090	0.116	0.132	0.150	0.160	0.176	0.184	0.200	0.207	0.213	0.234	0.259	0.300	0.362	0.413
	pow .8	0.170	0.205	0.229	0.249	0.265	0.282	0.295	0.309	0.320	0.330	0.351	0.378	0.418	0.480	0.528
	1% .05	0.123	0.162	0.194	0.222	0.248	0.272	0.294	0.314	0.333	0.350	0.383	0.425	0.484	0.570	0.631
	1% .01	0.196	0.234	0.267	0.296	0.321	0.345	0.366	0.386	0.404	0.421	0.452	0.493	0.547	0.626	0.681
	pow .5	0.121	0.140	0.154	0.169	0.180	0.188	0.203	0.210	0.216	0.231	0.251	0.275	0.305	0.365	0.415
	pow .8	0.206	0.231	0.251	0.268	0.283	0.296	0.310	0.320	0.331	0.343	0.363	0.389	0.425	0.485	0.532
	5% .05	0.207	0.231	0.255	0.277	0.298	0.317	0.335	0.352	0.369	0.384	0.413	0.450	0.503	0.583	0.640
	5% .01	0.288	0.312	0.334	0.355	0.375	0.393	0.410	0.426	0.442	0.456	0.483	0.518	0.566	0.639	0.690
	pow .5	0.200	0.211	0.219	0.231	0.237	0.243	0.256	0.262	0.266	0.279	0.288	0.309	0.344	0.398	0.444
	pow .8	0.294	0.306	0.317	0.329	0.339	0.348	0.359	0.367	0.375	0.386	0.400	0.421	0.455	0.509	0.551
50	nil .05	0.075	0.113	0.143	0.170	0.194	0.215	0.235	0.254	0.272	0.288	0.319	0.359	0.416	0.503	0.566
	nil .01	0.125	0.168	0.201	0.229	0.254	0.277	0.297	0.316	0.334	0.350	0.381	0.420	0.475	0.557	0.616
	pow .5	0.073	0.094	0.107	0.121	0.130	0.143	0.150	0.156	0.169	0.174	0.192	0.214	0.240	0.296	0.334
	pow .8	0.140	0.169	0.189	0.206	0.220	0.234	0.244	0.254	0.266	0.275	0.293	0.317	0.350	0.407	0.450
	1% .05	0.106	0.136	0.163	0.187	0.209	0.229	0.248	0.266	0.282	0.298	0.328	0.367	0.422	0.507	0.569
	1% .01	0.167	0.198	0.225	0.250	0.272	0.292	0.311	0.329	0.346	0.362	0.390	0.428	0.481	0.561	0.619
	pow .5	0.103	0.118	0.128	0.136	0.149	0.156	0.168	0.174	0.179	0.192	0.200	0.220	0.245	0.299	0.348
	pow .8	0.177	0.196	0.212	0.225	0.238	0.249	0.261	0.270	0.278	0.289	0.304	0.326	0.358	0.412	0.457
	5% .05	0.185	0.206	0.225	0.243	0.261	0.278	0.293	0.308	0.322	0.336	0.361	0.396	0.446	0.523	0.582
	5% .01	0.257	0.277	0.295	0.313	0.330	0.346	0.361	0.375	0.389	0.402	0.426	0.459	0.505	0.578	0.631
	pow .5	0.182	0.187	0.197	0.202	0.207	0.211	0.222	0.226	0.229	0.240	0.247	0.264	0.286	0.334	0.378
	pow .8	0.263	0.272	0.282	0.290	0.298	0.305	0.314	0.321	0.327	0.336	0.347	0.366	0.392	0.440	0.481

One-Stop PV Table

		df_{hyp}														
df_{err}		1	2	3	4	5	6	7	8	9	10	12	15	20	30	40
60	nil .05	0.062	0.095	0.121	0.144	0.165	0.184	0.202	0.218	0.234	0.249	0.277	0.314	0.368	0.452	0.515
	nil .01	0.105	0.142	0.171	0.196	0.218	0.238	0.256	0.273	0.290	0.305	0.333	0.370	0.423	0.503	0.563
	pow .5	0.061	0.079	0.090	0.102	0.109	0.115	0.126	0.131	0.135	0.147	0.155	0.173	0.205	0.245	0.280
	pow .8	0.118	0.144	0.161	0.176	0.187	0.198	0.209	0.217	0.225	0.235	0.249	0.270	0.302	0.350	0.390
	1% .05	0.094	0.119	0.141	0.162	0.181	0.198	0.215	0.231	0.246	0.260	0.287	0.323	0.375	0.456	0.519
	1% .01	0.147	0.173	0.196	0.217	0.237	0.255	0.272	0.288	0.303	0.317	0.344	0.379	0.430	0.508	0.567
	pow .5	0.091	0.103	0.111	0.117	0.128	0.133	0.138	0.149	0.153	0.157	0.171	0.189	0.211	0.260	0.294
	pow .8	0.156	0.172	0.185	0.196	0.207	0.216	0.224	0.234	0.241	0.248	0.263	0.283	0.310	0.360	0.399
	5% .05	0.172	0.188	0.204	0.220	0.235	0.249	0.263	0.276	0.288	0.301	0.324	0.355	0.402	0.476	0.534
	5% .01	0.235	0.251	0.267	0.282	0.297	0.311	0.324	0.337	0.349	0.361	0.383	0.413	0.458	0.528	0.582
	pow .5	0.166	0.174	0.178	0.182	0.190	0.194	0.197	0.200	0.210	0.213	0.218	0.233	0.251	0.295	0.326
	pow .8	0.241	0.249	0.256	0.262	0.270	0.276	0.281	0.287	0.295	0.299	0.309	0.325	0.348	0.391	0.425
70	nil .05	0.054	0.082	0.105	0.125	0.143	0.160	0.176	0.191	0.206	0.219	0.245	0.279	0.329	0.410	0.472
	nil .01	0.091	0.123	0.149	0.171	0.190	0.208	0.225	0.241	0.256	0.270	0.296	0.331	0.381	0.459	0.519
	pow .5	0.053	0.068	0.077	0.088	0.094	0.099	0.109	0.114	0.117	0.128	0.134	0.150	0.170	0.205	0.249
	pow .8	0.103	0.125	0.140	0.153	0.163	0.172	0.182	0.190	0.197	0.206	0.218	0.236	0.262	0.306	0.346
	1% .05	0.086	0.106	0.125	0.143	0.160	0.176	0.190	0.205	0.218	0.231	0.255	0.289	0.337	0.415	0.476
	1% .01	0.132	0.154	0.174	0.193	0.210	0.226	0.242	0.256	0.270	0.283	0.308	0.341	0.389	0.465	0.523
	pow .5	0.083	0.092	0.099	0.108	0.113	0.117	0.121	0.131	0.134	0.137	0.150	0.166	0.185	0.220	0.251
	pow .8	0.142	0.154	0.164	0.175	0.183	0.191	0.198	0.207	0.213	0.219	0.232	0.250	0.274	0.316	0.351
	5% .05	0.160	0.175	0.189	0.202	0.215	0.228	0.240	0.251	0.263	0.274	0.295	0.324	0.367	0.437	0.494
	5% .01	0.218	0.232	0.246	0.259	0.272	0.285	0.296	0.308	0.319	0.330	0.350	0.378	0.419	0.487	0.540
	pow .5	0.157	0.160	0.163	0.171	0.174	0.176	0.179	0.187	0.190	0.192	0.203	0.209	0.225	0.255	0.293
	pow .8	0.224	0.230	0.236	0.242	0.247	0.252	0.257	0.263	0.268	0.272	0.283	0.294	0.314	0.350	0.384

										df_{hyp}						
df_{err}		1	2	3	4	5	6	7	8	9	10	12	15	20	30	40
80	nil .05	0.047	0.072	0.093	0.111	0.127	0.142	0.157	0.170	0.183	0.196	0.219	0.251	0.298	0.375	0.435
	nil .01	0.080	0.109	0.131	0.151	0.169	0.185	0.201	0.215	0.229	0.242	0.266	0.299	0.346	0.421	0.480
	pow .5	0.046	0.060	0.068	0.077	0.083	0.087	0.096	0.100	0.103	0.106	0.118	0.133	0.151	0.183	0.213
	pow .8	0.091	0.111	0.124	0.136	0.145	0.153	0.162	0.168	0.175	0.180	0.193	0.210	0.233	0.273	0.307
	1% .05	0.079	0.096	0.113	0.129	0.144	0.158	0.171	0.184	0.196	0.208	0.230	0.261	0.307	0.381	0.440
	1% .01	0.121	0.140	0.158	0.174	0.190	0.204	0.218	0.231	0.244	0.256	0.278	0.309	0.355	0.428	0.485
	pow .5	0.078	0.083	0.089	0.097	0.101	0.105	0.108	0.117	0.120	0.122	0.134	0.140	0.165	0.197	0.226
	pow .8	0.130	0.139	0.149	0.158	0.165	0.172	0.178	0.186	0.191	0.196	0.208	0.221	0.246	0.284	0.316
	5% .05	0.152	0.164	0.176	0.189	0.200	0.212	0.222	0.232	0.243	0.253	0.272	0.298	0.338	0.405	0.460
	5% .01	0.204	0.217	0.229	0.241	0.253	0.264	0.274	0.285	0.295	0.305	0.323	0.349	0.388	0.453	0.505
	pow .5	0.147	0.153	0.156	0.158	0.161	0.162	0.170	0.172	0.174	0.176	0.186	0.190	0.204	0.231	0.257
	pow .8	0.211	0.216	0.221	0.225	0.229	0.233	0.239	0.243	0.247	0.250	0.260	0.270	0.287	0.319	0.347
90	nil .05	0.042	0.064	0.083	0.099	0.114	0.128	0.141	0.154	0.166	0.177	0.199	0.228	0.272	0.345	0.404
	nil .01	0.071	0.097	0.118	0.136	0.152	0.167	0.181	0.194	0.207	0.219	0.242	0.272	0.317	0.390	0.447
	pow .5	0.041	0.053	0.060	0.069	0.074	0.078	0.086	0.089	0.092	0.094	0.106	0.119	0.135	0.165	0.193
	pow .8	0.081	0.099	0.111	0.122	0.130	0.137	0.145	0.151	0.157	0.162	0.174	0.189	0.210	0.246	0.277
	1% .05	0.072	0.089	0.104	0.118	0.131	0.144	0.156	0.168	0.179	0.190	0.210	0.239	0.281	0.352	0.409
	1% .01	0.111	0.129	0.144	0.159	0.173	0.186	0.199	0.211	0.223	0.234	0.255	0.284	0.326	0.397	0.453
	pow .5	0.071	0.077	0.081	0.088	0.092	0.095	0.103	0.106	0.108	0.110	0.121	0.126	0.150	0.179	0.195
	pow .8	0.119	0.128	0.136	0.144	0.151	0.156	0.164	0.169	0.174	0.178	0.189	0.201	0.223	0.258	0.284
	5% .05	0.144	0.156	0.167	0.178	0.188	0.198	0.208	0.217	0.227	0.235	0.253	0.277	0.315	0.378	0.431
	5% .01	0.193	0.205	0.216	0.226	0.237	0.247	0.257	0.266	0.275	0.284	0.301	0.326	0.362	0.424	0.474
	pow .5	0.142	0.144	0.146	0.148	0.154	0.156	0.158	0.160	0.161	0.169	0.171	0.182	0.195	0.221	0.236
	pow .8	0.200	0.204	0.208	0.212	0.216	0.220	0.223	0.227	0.230	0.235	0.241	0.253	0.269	0.298	0.320

One-Stop PV Table

										df$_{hyp}$						
df$_{err}$		1	2	3	4	5	6	7	8	9	10	12	15	20	30	40
100	nil .05	0.038	0.058	0.075	0.090	0.103	0.116	0.128	0.140	0.151	0.161	0.182	0.209	0.251	0.320	0.377
	nil .01	0.064	0.088	0.107	0.123	0.138	0.152	0.165	0.177	0.189	0.200	0.221	0.250	0.292	0.362	0.418
	pow .5	0.037	0.048	0.055	0.062	0.067	0.070	0.077	0.081	0.083	0.085	0.096	0.108	0.123	0.150	0.166
	pow .8	0.074	0.090	0.101	0.110	0.118	0.124	0.132	0.137	0.142	0.147	0.158	0.172	0.191	0.225	0.250
	1% .05	0.067	0.082	0.096	0.108	0.121	0.132	0.143	0.154	0.164	0.174	0.193	0.220	0.260	0.327	0.383
	1% .01	0.104	0.119	0.133	0.147	0.160	0.172	0.183	0.194	0.205	0.215	0.235	0.262	0.303	0.370	0.424
	pow .5	0.066	0.071	0.075	0.081	0.085	0.087	0.094	0.097	0.099	0.101	0.111	0.115	0.128	0.154	0.179
	pow .8	0.111	0.119	0.126	0.133	0.139	0.144	0.150	0.155	0.159	0.163	0.173	0.184	0.202	0.233	0.260
	5% .05	0.139	0.150	0.158	0.168	0.178	0.187	0.196	0.204	0.213	0.221	0.237	0.260	0.295	0.355	0.406
	5% .01	0.184	0.195	0.205	0.214	0.224	0.233	0.242	0.250	0.259	0.267	0.283	0.306	0.340	0.399	0.448
	pow .5	0.135	0.137	0.142	0.144	0.145	0.147	0.148	0.155	0.156	0.157	0.160	0.169	0.181	0.204	0.218
	pow .8	0.190	0.194	0.198	0.201	0.204	0.208	0.211	0.215	0.218	0.221	0.226	0.236	0.250	0.277	0.297
120	nil .05	0.032	0.049	0.063	0.075	0.087	0.098	0.108	0.118	0.128	0.137	0.155	0.179	0.216	0.279	0.332
	nil .01	0.054	0.074	0.090	0.104	0.117	0.129	0.140	0.151	0.161	0.171	0.189	0.215	0.253	0.317	0.370
	pow .5	0.031	0.040	0.046	0.049	0.056	0.059	0.061	0.068	0.070	0.072	0.080	0.084	0.096	0.119	0.142
	pow .8	0.062	0.076	0.085	0.092	0.099	0.105	0.110	0.116	0.120	0.124	0.133	0.143	0.159	0.188	0.212
	1% .05	0.061	0.073	0.084	0.095	0.105	0.115	0.124	0.133	0.142	0.151	0.167	0.191	0.226	0.287	0.339
	1% .01	0.092	0.105	0.117	0.128	0.139	0.149	0.159	0.169	0.178	0.187	0.204	0.228	0.264	0.326	0.377
	pow .5	0.059	0.063	0.068	0.071	0.073	0.075	0.081	0.083	0.085	0.086	0.095	0.098	0.109	0.132	0.154
	pow .8	0.098	0.105	0.111	0.116	0.121	0.125	0.130	0.134	0.137	0.141	0.149	0.158	0.173	0.199	0.223
	5% .05	0.130	0.138	0.147	0.154	0.162	0.170	0.177	0.185	0.192	0.199	0.213	0.233	0.264	0.318	0.365
	5% .01	0.170	0.179	0.187	0.195	0.203	0.211	0.219	0.226	0.233	0.241	0.254	0.274	0.305	0.358	0.404
	pow .5	0.126	0.128	0.129	0.134	0.135	0.136	0.137	0.138	0.144	0.145	0.147	0.155	0.159	0.178	0.198
	pow .8	0.175	0.178	0.181	0.184	0.187	0.189	0.192	0.194	0.198	0.200	0.204	0.213	0.222	0.244	0.265

One-Stop PV Table

									df_{hyp}						
df_{err}		1	2	3	4	5	6	7	8	9	10	12	15	20	30	40
150	nil .05	0.025	0.039	0.051	0.061	0.070	0.079	0.088	0.096	0.104	0.112	0.127	0.148	0.179	0.235	0.282
	nil .01	0.043	0.060	0.073	0.084	0.095	0.105	0.114	0.123	0.132	0.140	0.156	0.178	0.211	0.268	0.315
	pow .5	0.025	0.032	0.037	0.040	0.045	0.047	0.049	0.054	0.056	0.058	0.065	0.068	0.078	0.097	0.108
	pow .8	0.050	0.061	0.069	0.075	0.080	0.085	0.089	0.094	0.097	0.101	0.108	0.116	0.129	0.153	0.170
	1% .05	0.054	0.063	0.072	0.080	0.089	0.096	0.104	0.112	0.119	0.126	0.140	0.160	0.190	0.243	0.289
	1% .01	0.080	0.090	0.099	0.108	0.117	0.126	0.134	0.142	0.149	0.157	0.171	0.192	0.223	0.277	0.323
	pow .5	0.052	0.056	0.058	0.060	0.062	0.066	0.068	0.069	0.071	0.071	0.078	0.081	0.090	0.108	0.119
	pow .8	0.086	0.090	0.094	0.098	0.102	0.106	0.109	0.112	0.115	0.117	0.124	0.131	0.143	0.165	0.181
	5% .05	0.120	0.126	0.132	0.139	0.146	0.152	0.158	0.165	0.170	0.176	0.187	0.204	0.230	0.277	0.319
	5% .01	0.155	0.162	0.168	0.175	0.181	0.188	0.194	0.200	0.206	0.212	0.224	0.240	0.267	0.313	0.354
	pow .5	0.116	0.120	0.121	0.122	0.123	0.123	0.124	0.125	0.130	0.130	0.132	0.139	0.141	0.157	0.167
	pow .8	0.160	0.163	0.164	0.166	0.168	0.169	0.171	0.173	0.176	0.178	0.181	0.188	0.194	0.212	0.226
200	nil .05	0.019	0.030	0.038	0.046	0.053	0.060	0.067	0.073	0.080	0.086	0.097	0.114	0.139	0.185	0.225
	nil .01	0.033	0.045	0.055	0.064	0.072	0.080	0.087	0.094	0.101	0.108	0.120	0.138	0.165	0.212	0.253
	pow .5	0.019	0.024	0.028	0.030	0.034	0.036	0.037	0.041	0.042	0.043	0.045	0.051	0.059	0.068	0.083
	pow .8	0.038	0.046	0.052	0.057	0.061	0.064	0.068	0.071	0.074	0.076	0.081	0.088	0.098	0.114	0.130
	1% .05	0.046	0.053	0.060	0.065	0.072	0.078	0.084	0.089	0.095	0.100	0.111	0.127	0.151	0.195	0.234
	1% .01	0.067	0.074	0.081	0.088	0.095	0.101	0.107	0.113	0.119	0.125	0.136	0.153	0.178	0.223	0.262
	pow .5	0.045	0.047	0.048	0.051	0.052	0.053	0.054	0.055	0.059	0.060	0.062	0.067	0.070	0.084	0.092
	pow .8	0.071	0.075	0.078	0.081	0.083	0.085	0.088	0.090	0.093	0.095	0.098	0.105	0.112	0.128	0.141
	5% .05	0.109	0.114	0.118	0.123	0.128	0.133	0.138	0.142	0.147	0.151	0.160	0.173	0.194	0.232	0.267
	5% .01	0.138	0.143	0.148	0.153	0.158	0.163	0.168	0.172	0.177	0.182	0.191	0.204	0.225	0.263	0.298
	pow .5	0.106	0.106	0.109	0.110	0.110	0.111	0.111	0.112	0.112	0.112	0.117	0.118	0.124	0.132	0.139
	pow .8	0.142	0.143	0.145	0.146	0.148	0.149	0.150	0.151	0.152	0.153	0.157	0.160	0.167	0.177	0.187

One-Stop PV Table

									df_{hyp}							
df_{err}		1	2	3	4	5	6	7	8	9	10	12	15	20	30	40
300	nil .05	0.013	0.020	0.026	0.031	0.036	0.041	0.045	0.050	0.054	0.058	0.067	0.078	0.097	0.130	0.160
	nil .01	0.022	0.030	0.037	0.043	0.049	0.054	0.059	0.064	0.069	0.073	0.082	0.095	0.114	0.150	0.181
	pow .5	0.013	0.016	0.018	0.020	0.023	0.024	0.025	0.025	0.028	0.029	0.030	0.035	0.040	0.046	0.052
	pow .8	0.025	0.031	0.035	0.038	0.041	0.044	0.046	0.047	0.050	0.052	0.055	0.060	0.066	0.077	0.086
	1% .05	0.037	0.042	0.046	0.050	0.054	0.058	0.062	0.066	0.070	0.074	0.081	0.092	0.109	0.141	0.170
	1% .01	0.053	0.058	0.062	0.067	0.071	0.075	0.079	0.084	0.088	0.092	0.100	0.111	0.129	0.162	0.192
	pow .5	0.036	0.037	0.038	0.040	0.041	0.042	0.042	0.043	0.043	0.046	0.047	0.048	0.053	0.058	0.064
	pow .8	0.056	0.058	0.060	0.062	0.063	0.064	0.066	0.067	0.068	0.070	0.073	0.076	0.081	0.091	0.099
	5% .05	0.097	0.100	0.103	0.107	0.109	0.112	0.115	0.119	0.122	0.125	0.131	0.140	0.154	0.182	0.208
	5% .01	0.119	0.122	0.126	0.129	0.132	0.136	0.139	0.142	0.145	0.149	0.155	0.164	0.179	0.207	0.233
	pow .5	0.094	0.095	0.095	0.095	0.098	0.098	0.098	0.098	0.099	0.099	0.099	0.103	0.104	0.108	0.113
	pow .8	0.122	0.123	0.123	0.124	0.126	0.126	0.127	0.127	0.128	0.129	0.130	0.133	0.136	0.142	0.149
400	nil .05	0.010	0.015	0.019	0.023	0.027	0.031	0.034	0.038	0.041	0.044	0.051	0.060	0.074	0.100	0.124
	nil .01	0.016	0.023	0.028	0.033	0.037	0.041	0.045	0.049	0.052	0.056	0.063	0.072	0.088	0.116	0.141
	pow .5	0.009	0.012	0.014	0.015	0.017	0.018	0.019	0.019	0.021	0.022	0.023	0.026	0.030	0.035	0.039
	pow .8	0.019	0.024	0.026	0.029	0.031	0.033	0.034	0.036	0.038	0.039	0.041	0.045	0.050	0.058	0.065
	1% .05	0.033	0.036	0.039	0.042	0.045	0.048	0.051	0.054	0.057	0.060	0.065	0.073	0.087	0.112	0.135
	1% .01	0.045	0.049	0.052	0.055	0.059	0.062	0.065	0.068	0.071	0.074	0.080	0.089	0.103	0.129	0.153
	pow .5	0.032	0.033	0.033	0.034	0.034	0.036	0.036	0.036	0.037	0.037	0.040	0.040	0.044	0.049	0.053
	pow .8	0.048	0.049	0.050	0.051	0.052	0.054	0.054	0.055	0.056	0.057	0.059	0.061	0.066	0.073	0.079
	5% .05	0.089	0.092	0.094	0.097	0.099	0.101	0.104	0.106	0.109	0.111	0.115	0.122	0.133	0.155	0.175
	5% .01	0.108	0.111	0.113	0.116	0.118	0.121	0.123	0.126	0.128	0.130	0.135	0.142	0.154	0.175	0.196
	pow .5	0.088	0.088	0.089	0.089	0.089	0.089	0.089	0.089	0.089	0.092	0.092	0.092	0.093	0.096	0.100
	pow .8	0.111	0.111	0.112	0.112	0.113	0.113	0.113	0.114	0.114	0.116	0.116	0.118	0.119	0.124	0.128

One-Stop PV Table

									df$_{hyp}$						
df$_{err}$		1	2	3	4	5	6	7	8	9	10	12	15	20	30	40
500	nil .05	0.008	0.012	0.015	0.019	0.022	0.025	0.028	0.030	0.033	0.036	0.041	0.048	0.060	0.081	0.102
	nil .01	0.013	0.018	0.022	0.026	0.030	0.033	0.036	0.039	0.042	0.045	0.051	0.059	0.071	0.094	0.115
	pow .5	0.008	0.010	0.011	0.012	0.014	0.014	0.015	0.015	0.017	0.018	0.018	0.021	0.022	0.028	0.032
	pow .8	0.015	0.019	0.021	0.023	0.025	0.026	0.028	0.029	0.030	0.031	0.033	0.036	0.039	0.047	0.052
	1% .05	0.030	0.032	0.035	0.037	0.040	0.042	0.044	0.046	0.049	0.051	0.056	0.062	0.073	0.093	0.113
	1% .01	0.040	0.043	0.046	0.048	0.051	0.053	0.056	0.058	0.061	0.063	0.068	0.075	0.086	0.108	0.128
	pow .5	0.029	0.030	0.030	0.030	0.031	0.031	0.032	0.033	0.033	0.033	0.033	0.036	0.037	0.040	0.043
	pow .8	0.043	0.044	0.044	0.045	0.046	0.046	0.048	0.048	0.049	0.049	0.051	0.053	0.055	0.061	0.065
	5% .05	0.085	0.087	0.089	0.091	0.092	0.094	0.096	0.098	0.100	0.102	0.105	0.111	0.120	0.137	0.154
	5% .01	0.101	0.103	0.105	0.107	0.109	0.111	0.113	0.115	0.117	0.119	0.123	0.128	0.138	0.155	0.173
	pow .5	0.083	0.083	0.083	0.083	0.083	0.083	0.085	0.085	0.085	0.085	0.086	0.086	0.088	0.089	0.092
	pow .8	0.103	0.103	0.104	0.104	0.104	0.104	0.106	0.106	0.106	0.106	0.107	0.108	0.110	0.112	0.116
600	nil .05	0.006	0.010	0.013	0.016	0.018	0.021	0.023	0.025	0.028	0.030	0.034	0.040	0.050	0.069	0.086
	nil .01	0.011	0.015	0.019	0.022	0.025	0.028	0.030	0.033	0.035	0.038	0.042	0.049	0.060	0.080	0.098
	pow .5	0.006	0.008	0.009	0.010	0.011	0.012	0.012	0.013	0.014	0.015	0.015	0.017	0.018	0.023	0.026
	pow .8	0.013	0.016	0.018	0.019	0.021	0.022	0.023	0.024	0.025	0.026	0.028	0.030	0.033	0.039	0.043
	1% .05	0.027	0.029	0.032	0.034	0.036	0.038	0.040	0.041	0.043	0.045	0.049	0.055	0.063	0.081	0.097
	1% .01	0.037	0.039	0.041	0.044	0.046	0.048	0.050	0.052	0.054	0.056	0.060	0.066	0.075	0.093	0.110
	pow .5	0.026	0.027	0.028	0.028	0.028	0.028	0.029	0.030	0.030	0.030	0.031	0.031	0.033	0.036	0.039
	pow .8	0.039	0.040	0.040	0.041	0.041	0.042	0.042	0.043	0.044	0.044	0.045	0.046	0.049	0.053	0.057
	5% .05	0.081	0.083	0.084	0.086	0.087	0.089	0.091	0.092	0.094	0.096	0.099	0.103	0.111	0.125	0.140
	5% .01	0.096	0.098	0.100	0.101	0.103	0.104	0.106	0.108	0.109	0.111	0.114	0.119	0.127	0.142	0.156
	pow .5	0.080	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.083	0.083	0.086	0.086
	pow .8	0.098	0.098	0.099	0.099	0.099	0.099	0.099	0.100	0.100	0.100	0.100	0.102	0.103	0.106	0.107

One-Stop PV Table

		df_{hyp}														
df_{err}		1	2	3	4	5	6	7	8	9	10	12	15	20	30	40
1000	nil .05	0.004	0.006	0.008	0.009	0.011	0.012	0.014	0.015	0.017	0.018	0.021	0.024	0.031	0.042	0.053
	nil .01	0.007	0.009	0.011	0.013	0.015	0.017	0.018	0.020	0.021	0.023	0.026	0.030	0.037	0.049	0.061
	pow .5	0.004	0.005	0.006	0.006	0.007	0.007	0.007	0.008	0.009	0.009	0.009	0.010	0.011	0.014	0.016
	pow .8	0.008	0.010	0.011	0.012	0.013	0.013	0.014	0.014	0.015	0.016	0.017	0.018	0.020	0.023	0.026
	1% .05	0.023	0.024	0.025	0.026	0.028	0.029	0.030	0.031	0.032	0.033	0.035	0.039	0.044	0.055	0.065
	1% .01	0.030	0.031	0.032	0.033	0.034	0.036	0.037	0.038	0.039	0.040	0.043	0.046	0.052	0.063	0.074
	pow .5	0.022	0.023	0.023	0.023	0.023	0.023	0.023	0.024	0.024	0.024	0.024	0.024	0.026	0.028	0.028
	pow .8	0.031	0.031	0.031	0.032	0.032	0.032	0.032	0.033	0.033	0.033	0.034	0.034	0.036	0.038	0.040
	5% .05	0.073	0.074	0.075	0.076	0.077	0.078	0.079	0.080	0.081	0.082	0.084	0.087	0.091	0.100	0.109
	5% .01	0.085	0.086	0.087	0.087	0.088	0.089	0.090	0.091	0.092	0.093	0.095	0.098	0.103	0.112	0.121
	pow .5	0.073	0.073	0.073	0.073	0.073	0.073	0.073	0.073	0.073	0.073	0.073	0.074	0.074	0.075	0.076
	pow .8	0.086	0.086	0.086	0.086	0.086	0.086	0.086	0.086	0.087	0.087	0.087	0.088	0.088	0.089	0.090
000	nil .05	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.002	0.002	0.002	0.002	0.003	0.004	0.006
0	nil .01	0.001	0.001	0.001	0.001	0.002	0.002	0.002	0.002	0.002	0.002	0.003	0.003	0.004	0.005	0.006
	pow .5	0.000	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	pow .8	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.002	0.002	0.002	0.002	0.002	0.003
	1% .05	0.013	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.015	0.015	0.015	0.016	0.017
	1% .01	0.015	0.015	0.015	0.015	0.015	0.016	0.016	0.016	0.016	0.016	0.016	0.017	0.017	0.018	0.019
	pow .5	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013
	pow .8	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015
	5% .05	0.057	0.057	0.057	0.057	0.057	0.057	0.057	0.057	0.057	0.058	0.058	0.058	0.059	0.059	0.060
	5% .01	0.060	0.060	0.060	0.060	0.060	0.060	0.061	0.061	0.061	0.061	0.061	0.061	0.062	0.063	0.064
	pow .5	0.057	0.057	0.057	0.057	0.057	0.057	0.057	0.057	0.057	0.057	0.057	0.057	0.057	0.057	0.057
	pow .8	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060

This page intentionally left blank

Appendix D

df Needed for Power = .80 ($\alpha = .05$) in Tests of Traditional Null Hypothesis

df_{err} Needed for Power=.80 (α = .05) in Tests of Traditional Null Hypothesis

ES		df_{hyp}																
PV	d	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
0.01	0.20	775	952	1072	1165	1260	1331	1394	1451	1504	1580	1670	1825	1992	2302	2565	3027	4016
0.02	0.29	385	473	533	579	627	662	694	722	762	787	832	909	993	1176	1313	1513	2010
0.03	0.35	255	313	353	384	416	439	460	479	505	522	552	603	660	782	874	1008	1341
0.04	0.41	190	233	263	286	310	328	343	358	377	390	413	451	494	585	654	774	1031
0.05	0.46	151	186	209	228	247	261	273	285	300	310	329	359	402	466	522	618	825
0.06	0.51	125	154	173	189	204	216	227	236	249	257	273	298	333	388	434	514	687
0.07	0.55	106	131	148	161	174	184	193	204	212	220	233	255	285	331	371	440	601
0.08	0.59	92	114	128	140	152	160	168	178	185	191	203	222	248	289	324	384	525
0.09	0.63	81	100	113	124	134	142	149	157	164	169	179	196	220	256	287	341	466
0.10	0.67	73	90	101	110	120	127	133	141	146	152	161	176	197	230	258	312	419
0.11	0.70	66	81	91	101	108	115	120	127	132	137	148	159	178	208	238	283	388
0.12	0.74	60	74	83	92	99	104	110	116	121	125	135	145	163	190	218	259	355
0.13	0.77	55	68	76	84	90	96	101	106	111	115	124	133	150	178	200	238	327
0.14	0.81	50	62	70	78	83	88	94	98	102	106	114	123	138	165	185	220	302
0.15	0.84	47	58	65	72	77	82	87	91	95	98	106	115	129	153	172	205	286
0.16	0.87	43	54	61	67	72	76	81	85	88	92	99	107	120	143	161	192	268

ES		df_{hyp}																
PV	d	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
0.17	0.91	40	50	57	63	68	72	76	80	83	86	93	101	112	134	151	183	251
0.18	0.94	38	47	53	59	63	67	71	75	78	81	87	96	106	126	142	172	236
0.19	0.97	36	44	50	55	59	63	67	70	73	77	82	90	101	119	136	163	227
0.20	1.00	34	42	47	52	56	60	64	67	69	73	77	85	96	112	129	154	214
0.22	1.06	30	37	42	47	51	54	57	60	62	65	70	76	86	102	116	139	194
0.24	1.12	27	34	39	42	46	49	52	54	57	59	63	69	78	93	105	128	178
0.26	1.19	25	31	35	38	42	44	47	49	52	54	58	63	71	85	96	117	163
0.28	1.25	22	28	32	35	38	41	43	45	48	49	53	58	65	78	90	107	152
0.30	1.31	21	26	30	32	35	37	40	42	44	45	49	53	61	72	83	100	142
0.32	1.37	19	24	27	30	33	35	37	39	40	42	45	50	56	68	76	93	131
0.34	1.44	18	22	25	28	30	32	34	36	38	39	42	46	52	63	72	87	123
0.36	1.50	16	20	23	26	28	30	32	33	35	37	39	43	49	59	67	81	115
0.38	1.57	15	19	22	24	26	28	30	31	33	34	37	40	45	55	62	76	108
0.40	1.63	14	18	20	23	24	26	28	29	31	32	35	38	43	52	59	72	101
0.42	1.70	13	17	19	21	23	24	26	27	29	30	32	36	40	48	55	67	96
0.44	1.77	12	15	18	20	21	23	24	26	27	28	30	33	38	45	52	64	91
0.46	1.85	12	14	17	19	20	22	23	24	25	26	29	31	35	43	49	60	85
0.48	1.92	11	14	16	18	19	20	22	23	24	25	27	30	34	40	46	57	81
0.50	2.00	10	13	15	16	18	19	20	21	22	24	25	28	32	38	44	54	76
0.60	2.45	8	10	11	12	13	14	15	16	17	18	19	21	24	29	33	41	58
0.70	3.06	6	7	8	9	10	11	11	12	13	13	15	16	18	22	25	30	44

This page intentionally left blank

Appendix E

df Needed for Power = .80 (α = .05) in Tests of the Hypothesis That Treatments Account for 1% or Less of the Variance in Outcomes

df_{err} Needed for Power =.80 (α = .05) in Tests of the Hypothesis That Treatments Account for 1% or Less of the Variance in Outcomes

ES		df_{hyp}																
PV	d	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
.02	.29	3225	3242	3301	3266	3334	3349	3364	3429	3442	3454	3479	3570	3621	3900	4042	4379	5260
.03	.35	1058	1086	1104	1122	1139	1176	1185	1199	1212	1254	1271	1303	1377	1518	1615	1833	2299
.04	.41	573	590	607	623	650	658	670	683	694	716	736	779	836	920	993	1151	1458
.05	.46	373	389	405	422	434	445	457	472	483	492	509	541	586	652	728	833	1075
.06	.51	269	285	299	313	323	334	343	357	365	373	387	414	450	506	568	654	854
.07	.55	208	223	235	246	255	267	275	283	290	297	315	338	362	419	460	533	718
.08	.59	166	180	192	202	211	219	226	236	243	249	265	279	307	357	393	457	606
.09	.63	139	151	161	170	180	187	193	199	208	214	224	241	266	303	343	400	532
.10	.67	117	130	139	147	154	162	168	174	179	187	196	212	234	268	297	355	473
.11	.70	101	113	122	129	136	143	149	154	159	166	174	189	205	240	266	318	426
.12	.74	89	99	108	115	121	127	133	138	142	147	157	170	185	217	241	289	388
.13	.77	80	89	97	104	109	114	121	125	129	133	142	152	168	197	220	264	355
.14	.81	72	80	87	94	99	104	110	114	118	121	130	139	154	181	202	243	327
.15	.84	65	73	80	86	91	95	101	105	108	112	120	128	142	168	187	225	303

ES		df_{hyp}																
PV	d	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
.16	.87	59	67	73	79	84	88	93	97	100	103	111	119	132	156	174	209	283
.17	.91	54	61	68	73	77	81	85	90	93	96	103	110	123	145	162	195	269
.18	.94	49	57	63	68	72	76	79	83	86	89	96	103	115	136	152	183	252
.19	.97	45	53	58	63	67	71	74	78	81	84	90	97	108	127	144	172	238
.20	1.00	42	49	55	59	63	67	69	73	76	79	84	91	101	120	137	162	224
.22	1.06	37	43	48	52	56	59	62	65	68	70	75	81	91	107	123	146	204
.24	1.12	32	38	43	47	50	53	56	59	61	63	68	74	83	97	111	134	185
.26	1.19	29	34	38	42	45	48	51	53	55	57	61	67	75	90	101	122	169
.28	1.25	26	31	35	38	41	43	46	48	50	53	56	61	69	82	92	111	156
.30	1.31	24	28	32	35	37	40	42	44	46	48	51	56	63	75	86	103	144
.32	1.37	21	26	29	32	34	36	39	40	42	44	47	52	58	69	79	96	135
.34	1.44	20	24	27	30	32	34	36	37	39	41	44	48	54	64	73	89	125
.36	1.50	18	22	25	27	29	32	33	35	36	38	41	44	50	60	68	83	117
.38	1.57	17	20	23	25	27	29	31	32	34	35	38	41	47	56	64	78	110
.40	1.63	15	19	21	23	25	27	29	30	32	33	35	39	44	53	60	73	103
.42	1.70	14	18	20	22	24	25	27	28	30	31	33	36	41	49	56	69	98
.44	1.77	13	16	19	20	22	24	25	27	28	29	31	34	39	47	53	65	92
.46	1.85	12	15	17	19	21	22	24	25	26	27	29	32	37	44	50	61	87
.48	1.92	12	14	16	18	19	21	22	23	25	26	28	30	34	41	47	58	82
.50	2.00	11	13	15	17	18	20	21	22	23	24	26	29	32	39	45	54	77
.60	2.45	8	10	11	13	14	14	16	16	17	18	19	21	24	29	34	41	59
.70	3.06	6	7	8	9	10	11	12	12	13	13	14	16	18	22	25	31	44

This page intentionally left blank

Author Index

Subject Index