2
EDITION

APPLIED POWER ANALYSIS FOR THE BEHAVIORAL SCIENCES

CHRISTOPHER L. ABERSON

ROUTLEDGE

APPLIED POWER ANALYSIS FOR THE BEHAVIORAL SCIENCES

Applied Power Analysis for the Behavioral Sciences is a practical "how-to" guide to conducting statistical power analyses for psychology and related fields. The book provides a guide to conducting analyses that is appropriate for researchers and students, including those with limited quantitative backgrounds. With practical use in mind, the text provides detailed coverage of topics such as how to estimate expected effect sizes and power analyses for complex designs. The topical coverage of the text, an applied approach, in-depth coverage of popular statistical procedures, and a focus on conducting analyses using R make the text a unique contribution to the power literature. To facilitate application and usability, the text includes ready-to-use R code developed for the text. An accompanying R package called pwr2ppl (available at https://github.com/chrisaberson/pwr2ppl) provides tools for conducting power analyses across each topic covered in the text.

Christopher L. Aberson is Professor of Psychology at Humboldt State University. His research interests in social psychology include prejudice, racism, and attitudes toward affirmative action. His quantitative interests focus on statistical power. He serves as Editor-in-Chief for *Analyses of Social Issues and Public Policy*.

APPLIED POWER ANALYSIS FOR THE BEHAVIORAL SCIENCES

2nd Edition

Christopher L. Aberson

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

Second edition published 2019
by Routledge
52 Vanderbilt Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

First edition published by Routledge 2010

Library of Congress Cataloging-in-Publication Data
A catalog record has been requested for this book

ISBN: 978-1-138-04456-2 (hbk)
ISBN: 978-1-138-04459-3 (pbk)
ISBN: 978-1-315-17150-0 (ebk)

Typeset in Bembo
by Wearset Ltd, Boldon, Tyne and Wear

Visit the online resources at https://github.com/chrisaberson/pwr2ppl

CONTENTS

FIGURES

TABLES

PREFACE

Statistical power analyses differ in important ways from other statistical approaches. Most statistical analyses begin with existing data, subject the data to analysis, and then focus on interpretation of the results. Power analysis is different. Power analysis does not involve existing data. In fact, power analyses are only meaningful when conducted prior to data collection. In this manner, it is useful to think of power analysis as part of the hypothesis statement process. When stating a hypothesis, it is usually of the form of "Group X differs from Group Y" on our dependent measure. The statistical core of this statement is Group X and Group Y will differ. For power analysis, we go beyond this basic statement and specify how large a difference would be meaningful to detect between the two groups.

Another way power analysis differs from other statistical analyses is in terms of interpretation. For most statistical procedures, texts devote considerable time to interpretation of result or computer output. In contrast, the output for power analysis is simple and requires little interpretation or discussion. Generally, output provides a single value, the power for the test of interest. The interpretation of output for such analyses does not involve much interpretation aside from an evaluation of whether our study is sensitive enough to detect our effects of interest given a particular sample size.

This book also differs considerably from earlier texts on the topic (e.g., Cohen, 1988) in that I do not present power tables or formula for extrapolating between tabled values. Instead, most chapters present hand calculations to facilitate conceptual understanding but rely heavily on computer-generated analyses as the primary approaches for analyses. Given the computational tools available in R, approaches that involve reference to lengthy tables are no longer necessary.

Overview of the Book

Chapter 1 reviews significance testing, introduces power, and presents issues impacting power. Chapters 2 through 9 cover power analysis strategies for a variety of common designs. Chapter 2 (Chi square and proportions) and Chapter 3 (*t*-tests) also introduce issues such as noncentral distributions and provide examples of the types of decisions and considerations important to power analyses. Regardless of the technique of interest for your design, read Chapters 1–3 first. Chapter 4 covers power for correlations and for tests comparing correlations. Chapters 5 through 7 address ANOVA designs for between, within, and mixed models as well as Multivariate ANOVA. Chapter 8 covers multiple regression, comparisons of regression coefficients, and detecting multiple effects in the same study. Chapter 9 addresses covariate designs, regression interactions, logistic regression (LR), and mediation. Chapter 10 focuses on precision analysis for confidence intervals (CI) around mean difference, correlations, and effect sizes. Chapter 11 addresses a number of smaller topics such as how to report power analyses and how to increase power without increasing sample size. Chapters focusing on simpler analyses (e.g., *t*-test, between subjects ANOVA) present detailed formulae and calculation examples. However, chapters focusing on more complex topics (e.g., within subjects ANOVA, ANCOVA) present only computer-based analyses as calculation examples would extend several pages and do little to advance understanding.

What is New in this Edition?

The biggest change from the first to second edition involves statistical software. This edition uses R for all power calculations whereas the first edition used SPSS. The major additions to the topical coverage are expanded sections on power for detecting multiple effects in the same model (primarily in the multiple regression chapter), linear mixed model approaches for designs including within subjects factors, logistic regression, and mediation.

Formulae and Calculations

A major focus of this text is conducting analyses using R. However, understanding the basics of the calculations surrounding analyses is very important. To that end, whenever it is possible and not too complicated, I provide detailed calculations for sample analyses. Often these calculations involve several steps and multiple formulae. One of my points of contention with many statistical resources is that calculations are often not clearly detailed. That is, authors present formulae and then jump to the result without demonstrating the steps of the calculation. When I encounter this approach, it often takes some time to figure out what goes where and how the authors

derived values. For this reason, I provide detailed calculations, comment on what goes where, and how it got there. This approach is likely a bit more like an introductory than an advanced statistics text. However, the added detail makes the calculations easier to follow.

In many places throughout the text, I present formulae for population values. In practice, we rarely perform such calculations. I present population values to serve as a reminder that the calculations involved in power analyses generally involve a priori estimates of what the population looks like (e.g., estimates of the population effect size).

Approaches to Power

Several chapters provide three different approaches to the calculation of power. The first involves estimation of power. Estimation involves use of central rather than noncentral distributions. I debated inclusion of estimation techniques. On the one hand, estimation approaches enhance understanding of constructs through direct calculation of values. On the other hand, estimation does not always yield accurate power because it uses the wrong distribution. Ultimately, I included estimation procedures, as these techniques are excellent teaching tools. The values may not be completely accurate, but the conceptual piece is clearer with estimation demonstrations.

The next approach involves hand calculations with R to calculate power. Hand calculations provide accurate values for every estimate required for power analyses, except the power value itself. Hand calculations end at the noncentrality parameter. We then take that value to R for calculation of power. This is because power calculations require the use of noncentral distributions. Calculations based on noncentral distributions are not practical to complete by hand, as they involve numerous iterations. When completing hand calculations in several chapters, I include a single line of R code that handles the final step of the calculation.

The final approach involves use of R functions for all calculations (described in more detail below). I present these approaches in Chapters 2–10. This approach requires input of descriptive statistics and few calculations.

The pwr2ppl Companion Package

The first edition of this book presented complex SPSS Syntax approaches for conducting power analyses. Since that time, I have largely abandoned the SPSS environment. R is free and, in my view, infinitely more powerful than SPSS. I was able to add several new approaches to this book that could not be easily addressed in SPSS. I believe the approaches using R are considerably simpler than the SPSS materials presented in the first edition. Most approaches require a single line of code.

As a companion to this book, I compiled all of the function in this book in an R package called pwr2ppl. This is available from https://github.com/chrisaberson/pwr2ppl. To install the package (using the devtools package), simply type devtools::install_github(chrisaberson/pwr2ppl, dependencies = TRUE). I expect to continue to expand the pwr2ppl package after the publication of the text.

ACKNOWLEDGMENTS

The idea for this text came about following preparation for a workshop presentation on power in 2007. Jodie Ullman invited me to teach the workshop. Without this invitation, I would never have written this book. Dale Berger provided detailed and incisive comments on drafts of most of these chapters in the first edition that helped me to improve the text considerably. Thank you Dale for the extraordinary time you spent helping me with this project. Geoff Cumming allowed use of figures produced by his ESCI software as well as helpful comments on using the software.

As in the first edition, I want to thank my wife Nanda and son Ernesto for their love and support. Special thanks to our second and third editions, Paloma and Lucia, who arrived after the publication of the first edition of this book.

1

WHAT IS POWER?

Why is Power Important?

This chapter reviews null hypothesis significance (NHST) testing, introduces effect sizes and factors that influence power, discusses the importance of power in design, presents an introduction to noncentral distributions, addresses misconceptions about power, discusses typical levels of power in published work, examines strategies for determining an appropriate effect size for power analysis, critiques post hoc power analyses, and discusses typical levels of power used for design.

Review of Null Hypothesis Significance Testing

NHST focuses on conditional probabilities. The conditional probabilities used in NHST procedures address how likely it is to obtain an observed (i.e., sample) result given a specific assumption about the population. Formally, the assumption about the population is called the null hypothesis (e.g., the population mean is 0) and the observed result is what the sample produces (e.g., a sample mean of 10). Statistical tests such as z, χ^2, t, and Analysis of Variance (ANOVA) determine how likely the sample result or any result more distant from the null hypothesis would be if the null hypothesis were true. This probability is then compared to a set criterion. For example, if a result this far or farther from the null hypothesis would occur less than 5% of the time when the null is true, then we will reject the null. More formally, the criterion is termed a Type I or α error rate (5% corresponds to $\alpha = .05$).

Table 1.1, common to most introductory statistical texts, summarizes decisions about null hypotheses and compares them to what is true for the data ("reality"). Two errors exist. A Type I or α error reflects rejecting a true null hypothesis. Researchers control this probability by setting a value for it (e.g.,

use a two-tailed test with $\alpha = .05$). Type II or β errors reflect failure to reject a false null hypothesis. Controlling this probability is at the core of this book. Type II errors are far more difficult to control than Type I errors. Table 1.1 also represents correct decisions, either failing to reject a true null hypothesis or rejecting a false null. The probability of rejecting a false null hypothesis is power. As suggested by the title of this book, this topic receives considerable coverage throughout the text.

For power analysis, the focus is on situations for which the expectation is that the null hypothesis is false (see Chapter 10 for a discussion of approaches to "supporting" null hypotheses). Power analysis addresses the ability to reject the null hypothesis when it is false.

Effect Sizes and Their Interpretation

One of the most important statistics for power analysis is the effect size. Significance tests tell us only whether an effect is present. Effect sizes tell us how strong or weak the observed effect is.

Although researchers increasingly present effect size alongside NHST results, it is important to recognize that the term "effect size" refers to many different measures. The interpretation of an effect size is dependent on the specific effect

TABLE 1.1 Reality vs. Statistical Decisions

Reality			
		Null Hypothesis True	*Null Hypothesis False*
Research Decision	Fail to Reject Null	Correct failure to reject null $1-\alpha$	Type II or β error
	Reject Null	Type I or α error	Correct rejection of null $1-\beta$

TABLE 1.2 Measures of Effect Size, Their Use, and a Rough Guide to Interpretation

Effect Size	*Common Use/Presentation*	*Small*	*Medium*	*Large*
Φ (also known as V or w)	Omnibus effect for χ^2	0.10	0.30	0.50
h	Comparing proportions	0.20	0.50	0.80
d	Comparing two means	0.20	0.50	0.80
r	Correlation	0.10	0.30	0.50
q	Comparing two correlations	0.10	0.30	0.50
f	Omnibus effect for ANOVA/Regression	0.10	0.25	0.40
η^2	Omnibus effect for ANOVA	0.01	0.06	0.14
f^2	Omnibus effect for ANOVA/Regression	0.02	0.15	0.35
R^2	Omnibus effect for Regression	0.02	0.13	0.26

size statistic presented. For example, a value of 0.14 would be relatively small in discussing the d statistic but large when discussing η^2. For this reason, it is important to be explicit when presenting effect sizes. Always reference the value (d, r, η^2, etc.) rather than just noting "effect size."

Table 1.2 provides a brief summary of common effect size measures and definitions of small, medium, and large values for each (Cohen, 1992). Please note that the small, medium, and large labels facilitate comparison across effects. These values do not indicate the practical importance of effects.

What Influences Power?

I learned an acronym in graduate school that I use to teach about influences on power. That acronym is BEAN, standing for Beta (β), Effect Size (E), Alpha (α), and Sample Size (N). We can specify any three of these values and calculate the fourth. Power analysis often involves specifying α, effect size, and Beta to find sample size.

Power is $1-\beta$. As α becomes more liberal (e.g., moving from .01 to .05), power increases. As effect sizes increase (e.g., the mean is further from the null value relative to the standard deviation), power increases. As sample size rises, power increases.

Several figures represent the influence of effect size, α, and sample size on power. Figure 1.1 presents two distributions: the null and the alternative. The null distribution is the distribution specified in the null hypothesis and

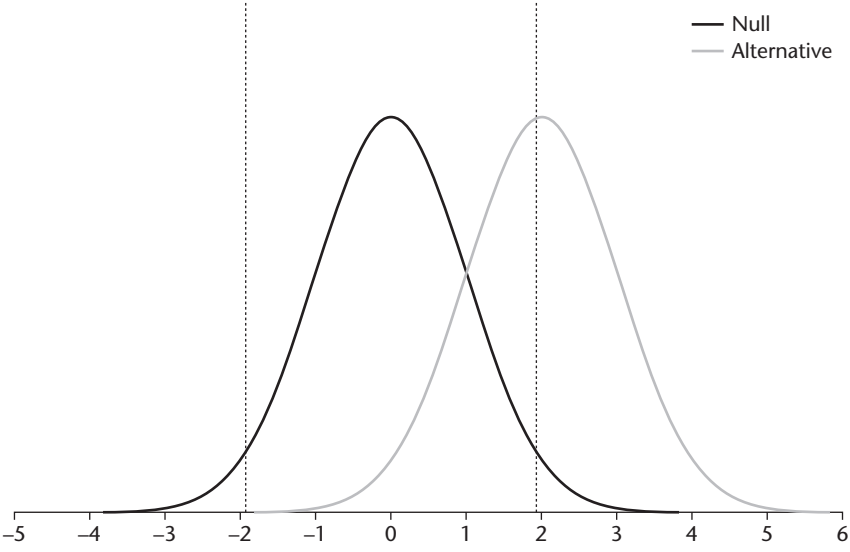

FIGURE 1.1 Null and Alternative Distributions for a Two-Tailed Test and $\alpha = .05$.

represented on the left hand side of the graph. For this example, the null hypothesis is that the population mean is zero. The null sampling distribution of the mean is centered on zero, reflecting this null hypothesis. The alternative sampling distribution, found on the right hand side of each graph, reflects the distribution of means from which we are actually sampling. Several additional figures follow and are useful for comparison with Figure 1.1. For simplicity, the population standard deviation and the null hypothesis remain constant for each figure. For each of the figures, the lines represent the $z_{critical}$ values.

A sample mean allows for rejection of the null hypothesis if it falls outside the critical values that we set based on the null distribution. The vertical lines in Figure 1.1 represent the critical values that cut off 2.5% in each tail of the null distribution (i.e., a two-tailed test with $\alpha = .05$). A little more than half of samples drawn from the alternative distribution fall above the upper critical value (the area to the right of the line near +2.0). Sample means that fall above the critical value allow for rejection of the null hypothesis. That area reflects the power of the test, about .50 in this example.

Now compare the power in Figure 1.1 to power in Figure 1.2. The difference between the situations represented in these two figures is that the effect size, represented in terms of the difference between the null and alternative means, is larger for Figure 1.2 than Figure 1.1 (recall that standard deviation is constant for both situations). The second figure shows that as the effect size increases, the distributions are further apart, and power increases because more of the alternative distribution falls in the rejection region.

Next, we consider the influence of α on power. Figures 1.1 and 1.2 presented a two-tailed test with $\alpha = .05$. Figure 1.3 reduces $\alpha = .01$. Notice the

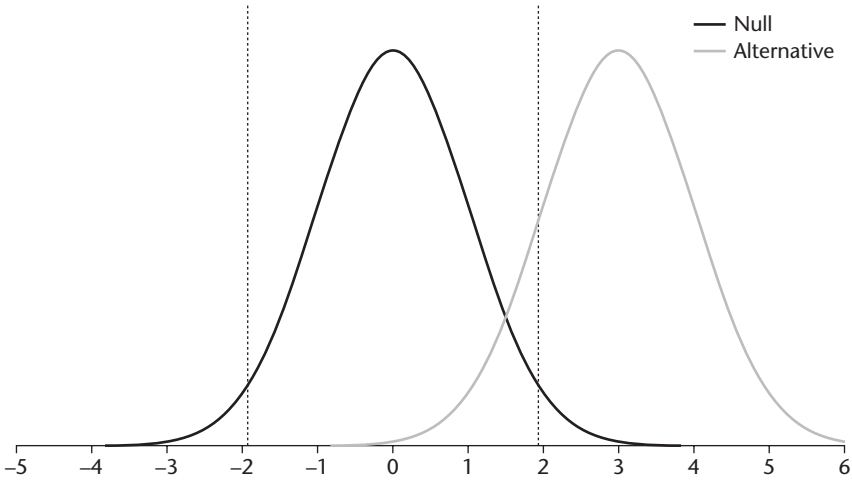

FIGURE 1.2 Null and Alternative Distributions for a Two-Tailed Test With Increased Effect Size and $\alpha = .05$.

change in the location of the vertical lines representing the critical values for rejection of the null hypothesis. Comparing Figures 1.1 and 1.3 shows that reducing α decreases power. The area in the alternative distribution that falls within the rejection region is smaller for Figure 1.3 than 1.1. Smaller values for α make it more difficult to reject the null hypothesis. When it is more difficult to reject the null hypothesis, power decreases.

Figure 1.4 demonstrates the influence of a larger sample size on power. This figure presents distributions that are less disperse than those in Figure 1.1. For

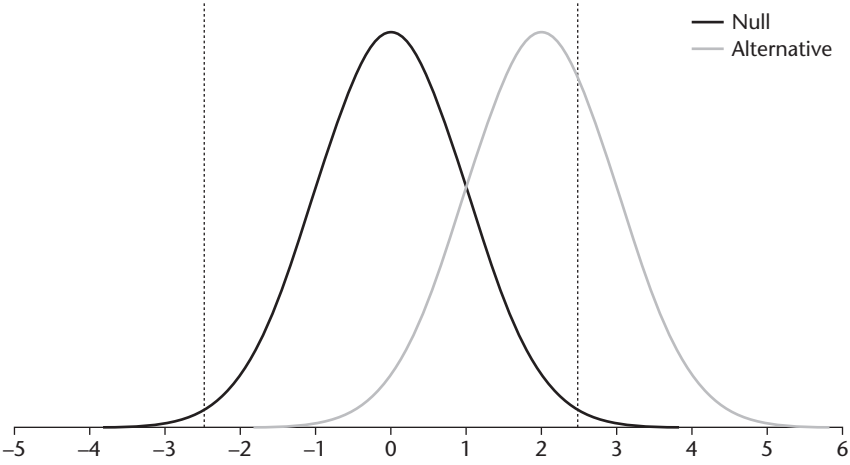

FIGURE 1.3 Null and Alternative Distributions for a Two-Tailed Test with $\alpha = .01$.

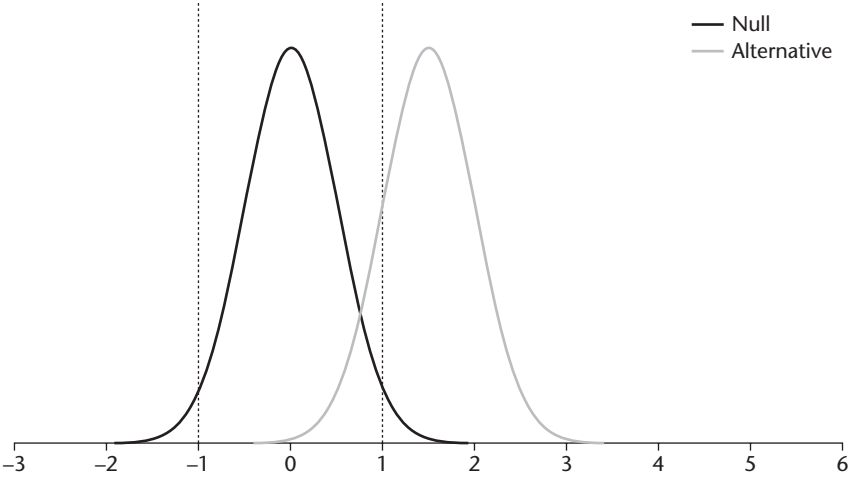

FIGURE 1.4 Null and Alternative Distributions for a Two-Tailed Test and $\alpha = .05$ with a Large Sample.

this figure, the *x*-axis represents raw scores rather than *z*-values. Recall from the Central Limit Theorem that the dispersion of a distribution of sample means (standard error of the mean) is a function of the standard deviation in the population and the sample size. Specifically, this is the standard deviation divided by the square root of sample size. As sample size rises, dispersion decreases. As seen in Figure 1.4 (assuming that the difference between the means of the distributions are the same for Figures 1.1 and 1.4), the reduced dispersion that results from larger samples increases power.

For an interactive tutorial on the topic, see the WISE (Web Interface for Statistics Education) home page at wise.cgu.edu. The web page includes a detailed interactive applet and tutorial on power analysis (see Aberson, Berger, Healy, & Romero, 2002 for a description and evaluation of the tutorial assignment). Chapters 2 and 3, particularly the material relevant to Figures 2.1 and 3.1, provide descriptions useful for power calculations.

Central and Noncentral Distributions

The examples presented in the preceding section use the normal distribution. In practice, these tests are less common in most fields than those conducted using *t*, *F*, or χ^2. Power analyses become more complex when using these distributions. Power calculations are based on the alternative distribution. When we use a *z*-test with a normally distributed null distribution, the alternative distribution is also normally distributed no matter the size of the effect. Distributions of this type are termed *central distributions*. However, when we deal with other tests, the alternative distribution takes on different shapes that vary with effect size. These are termed *noncentral distributions*. When conducting tests such as *t*, *F*, or χ^2 we actually deal with both central and noncentral distributions. The null distribution is a central distribution and the alternative distribution is a noncentral distribution.

Central and noncentral distributions differ in important ways. Degrees of freedom are the only influence on the shape of central distributions. The null distribution when using *t*, *F*, or χ^2 is a central distribution. Since the null hypothesis specifies no effect, these distributions correspond to situations for which the effect size is zero (i.e., effect size is constant). Thus, the shape of the null distribution varies only with degrees of freedom. For any specific value for degrees of freedom, there is just one distribution used to compute probabilities. For example, if we have a *t*-distribution with $df = 50$, we can calculate the area at or above $t = 2.9$ (or any other value).

Degrees of freedom and effect size define the shape of noncentral distributions. For any specific value for degrees of freedom, there are infinite possible effect sizes. Since there are infinite possible values for effect size for each of the infinite possibilities for degrees of freedom, there are simply too many possible combinations for construction of tables that allow for simple calculation of

probabilities. Suffice it to say, calculations of probabilities associated with these distributions are far more complicated than for central distributions. The text provides several R functions and code examples for performing these calculations.

Figure 1.5[1] demonstrates differences between central and noncentral distributions using an example of a *t*-distribution with 10 degrees of freedom. The null distribution on the left is symmetrical. Recall that the null distribution is a central distribution. The distribution on the right, however, is nonsymmetrical. This is the noncentral *t*-distribution. This is the distribution used to calculate power. The noncentral *t* represents the actual population distribution for *t* from which we are sampling. As in previous examples, the critical value is defined in relation to the null distribution, but calculation of power focuses on where the critical value falls in relation to the alternative distribution (on the right). On this figure, the vertical line slightly to the left of 2.30 on the *x*-axis represents the $t_{critical}$ value. Because the shape of the noncentral *t*-distribution depends on both the degrees of freedom and the effect size, there is no simple table such as we find in the back of statistics texts for the central *t*-distribution. Calculation of area based on the noncentral distribution is much more difficult.

Noncentral distributions sometimes look similar to central distributions. Figure 1.6[2] demonstrates a situation with 100 degrees of freedom. Again, the vertical line slightly to the left of 2.30 on the *x*-axis represents the *t*-critical

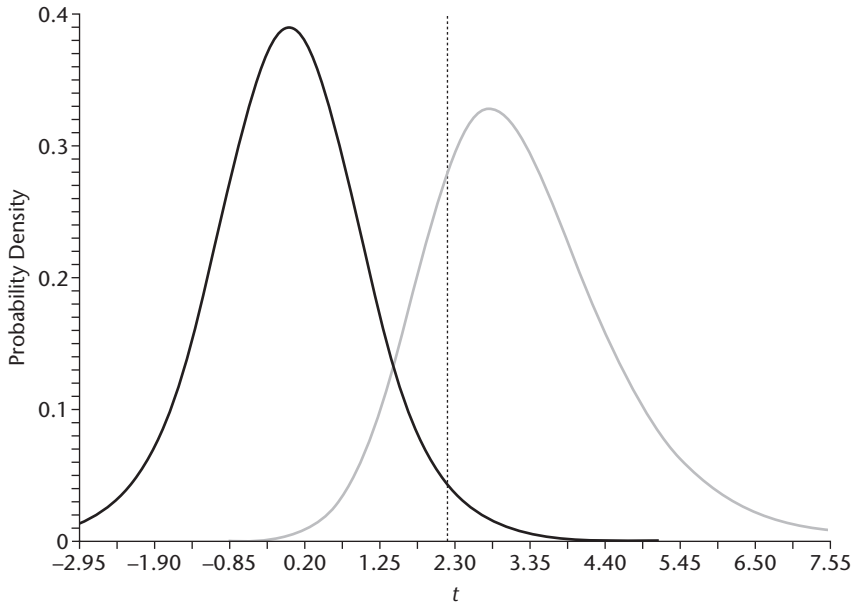

FIGURE 1.5 Central vs. Noncentral *t*-distributions ($df = 10$).

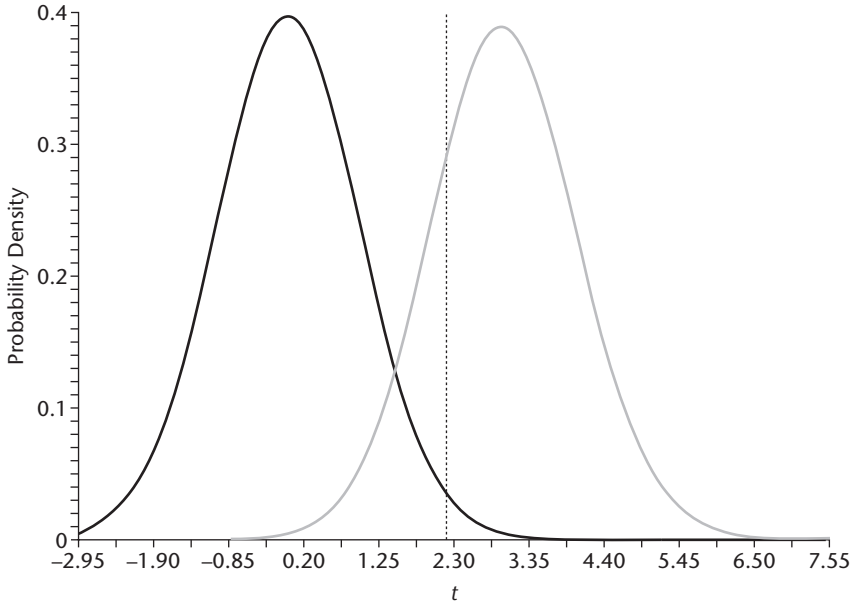

FIGURE 1.6 Central vs. Noncentral *t*-distributions ($df = 100$).

value. Here, the central and noncentral distributions are similar. As sample size increases, central and noncentral distribution shapes begin to converge. In situations like this, approximations of power using central distributions produce reasonably accurate estimates. However, given the availability of computer protocols that are presented in this book, approximation is a poor strategy for analysis (but see the discussion of the value of approximation in this chapter).

Misconceptions about Power

Students and researchers often misunderstand factors relating to statistical power. One common misunderstanding is the relationship between Type II and Type I errors. Given a design with a 5% Type I error rate, researchers often predict the rate of Type II errors also to be 5% (Hunter, 1997; Nickerson, 2000). The probability of a Type II error is generally much greater than 5%, and in a given study, the probability of a Type II error is inversely related to the Type I error rate. In practice, a 5% Type II error rate is small. A commonly recommended goal for power is 80%. This corresponds to a Type II error rate of 20%.

Another misconception is the belief that failure to reject the null hypothesis is sufficient evidence that the null hypothesis is true (Nickerson, 2000). Even when power is 80% and we have specified the population effect size accurately, there still exists a 20% chance of making a Type II error. Compare this rate to

the commonly used Type I error rate of 5%. The Type I rate suggests that falsely rejecting a true hypothesis occurs 1 time out of 20. The Type II rate suggests that failing to reject a false null hypothesis happens one time out of every five samples. Thus, when researchers make claims about supporting the null hypothesis based on failing to reject the null hypothesis, these claims usually provide little statistical evidence.

Empirical Reviews of Power

Surveys of the power in published work in the behavioral sciences indicate that samples generally do not provide adequate power for detecting small and medium size effects. Power surveys across fields such as abnormal psychology (e.g., Cohen, 1962; Sedlmeier & Gigerenzer, 1989); management (Cashen & Geiger, 2004; Mone, Mueller, & Mauland, 1996); rehabilitation counseling (Kosciulek & Szymanski, 1993); psychiatry (Brown & Hale, 1992); behavioral ecology and animal behavior (Jennions & Møller, 2003); adult education (West, 1985); consulting and clinical psychology (Rossi, 1990); neuroscience (Button et al., 2013); and social and personality (Fraley & Vazire, 2014) all suggest that low power is common and has been low for some time. For small effects, these surveys report typical power levels of .13 to .26, indicating that power to detect small effects is generally low. Power to detect medium effects was higher with reported ranges from .39 to .98. Not surprisingly, power for detecting large effects was best, ranging from .77 to .99.

Notably, in a review of articles published over a 1-year period in major health psychology outlets, power was consistently higher than reported in the reviews just mentioned (Maddock & Rossi, 2001). This finding existed across small (Power = .36), medium (.77), and large (.92) effect sizes. For small and medium effects, studies reporting federal funding were more powerful (.41 for funded vs. .28 for not funded for small effects and .80 for funded vs. .71 for not funded for medium effects). Federally funded projects often require power analyses as part of the grant submission so requiring power analyses appears to promote more sensitive designs.

The relatively poor statistical power observed in published work suggests that either researchers do not conduct power analyses or if they do conduct power analyses, they assume large effects. However, the Maddock and Rossi (2001) study found that when required to conduct power analyses, researchers designed studies with more power. Broadly speaking, this suggests that when left to their own devices, researchers tend to design underpowered studies.

Another possible reason for low power in published work reflects a fundamental misunderstanding of how to address issues of effect size for sample size planning. Most power analyses begin with an effect size estimate. The researcher determines sample size based on effect size, the desired level of power, and Type I error. In my experience, estimation of the effect size is problematic to the point of

being arbitrary (see also Lipsey, 1990). There are several reasons for this. Researchers often use standard effect size estimates (i.e., small, medium, and large) without reference to the size of an effect that would be practically important. In consulting on power analyses, it is my sense that many researchers think that they are being conservative if they choose a medium effect size. Compounding these problems are failures to understand typical effect sizes in the field of inquiry. Many fields in the behavioral sciences deal with small effects, clearly calling into question the "medium effect size as conservative approach."

Another issue is that researchers sometimes begin with a conservative effect size estimate but then, discouraged by large sample size requirements for adequate power, increase their effect size to reduce sample size. A related problem occurs when researchers begin with an idea of the sample size they want and design backward to find the effect size they plan to detect.

Consequences of Underpowered Studies

A typical underpowered study has a relatively low probability of detecting a statistically significant effect compared to a study with sufficient power. Historically, within the behavioral sciences, significant results are far more likely to get published than nonsignificant results (often discussed as publication bias or the file-drawer problem). At first blush, this appears to be a problem for the researchers conducting the study. No significant findings leads to unpublishable work. Unfortunately, there is also a serious issue regarding underpowered studies that do find significant results.

When a study is underpowered, statistically significant results occur only when effect sizes are larger than the effect size in the population. As an example, a study with Power = .20 and a population effect size $d = 0.50$ requires a sample effect size of $d = 0.89$ or larger for a statistically significant result. In contrast, when Power = .80, the effect size required for a significant result is $d = 0.35$ or higher. Statistically significant results from the underpowered study reflects a considerable overestimation of the true population effect whereas those from well-powered studies provide a range of values both above and below the actual population effect size. As power appears low for many fields in the behavioral sciences and publication bias remains the norm at many outlets, this suggests that published literature provides a skewed view of true population effects.

Low power and skewed views of population effects produce a body of research that does not stand up to replication (e.g., Open Science Collaboration, 2015). Although there are many factors driving failures to replicate, low statistical power is among the most prominent. So much so that recent editorials highlight issues with power. For example, Stephan Lindsay, in an editorial in *Psychological Science* stated that based on power analyses in submitted manuscripts, "my impression is that many authors … have but a shaky grasp on that concept" (2015, p. 1828). Consistent with this perception, a study that surveyed published research psychologists

revealed widespread overestimation of power and other poor intuitions regarding power (Bakker, Hartgerink, Wicherts, & van der Maas, 2016).

Echoing the call for a focus on power, Simine Vazire, editor of *Social and Personality Psychological Science* wrote "among all the proposed reforms … I am most interested in increasing the statistical power of our published studies" (2016, p. 4). Several other outlets such as *Journal of Personality and Social Psychology* (American Psychological Association, n.d.a), *Emotion* (American Psychological Association, n.d.b), *Journal of Research in Personality* (Lucas & Donnellan, 2013), *Social Psychology* (Unkelbach, 2016), and *Nature* (Nature Publishing Group, 2017) demonstrate a substantially increased focus justifying sample size.

The *American Psychological Association Publication Manual* clearly states that authors should "[s]tate how this intended sample size was determined (e.g., analysis of power…;" 2010, p. 30). These recommendations date back several decades (see Wilkinson & Task Force on Statistical Inference, 1999). Despite these calls for justifications of sample size via power analysis, it does not appear that most authors and outlets adhered to such recommendations. The editorials cited previously and the actions of other editors and journals suggest that one of the major consequences for the behavioral sciences of underpowered studies is an increased focus on providing clear and meaningful justifications of sample size (i.e., power analysis) as a requirement for publication.

Overview of Approaches to Determining Effect Size for Power Analysis

Perhaps the most difficult requirement of power analysis is estimation of effect size. Designing for too small an effect wastes resources through collection of more data than needed. Designing for too large an effect does not achieve the desired level of power. Unfortunately, determining an effect size for power analysis is not always easy. Good estimates of effect size require careful consideration. This section reviews strategies for determining effect size for power analyses and presents critiques of each approach.

Perhaps the most important point in this section is that the effect size you choose influences the outcome of the power analysis more than any other decision. Do not take this choice lightly. Good power analyses start with informed choices about effects. The more time, effort, and thought put into this estimate, the better the analysis.

Determination of Expectations for Small, Medium, or Large Effects

Use of arbitrary effect size estimates (e.g., small, medium, large) is a bad approach. Sometimes this approach is the most effective or useful approach available, but premature use takes consideration away from important issues such as whether effects are meaningful (discussed in greater detail in the next

section), the raw differences we wish to detect, and the precision of measurement. Lenth (2000) gives a useful demonstration of these issues in a test for a medium effect. One example involves a between subjects test that detects relatively a roughly 1 mm difference between groups using an instrument with a relatively large standard deviation (1.9 mm). The second test involves a paired test using a more precise instrument ($SD = 0.7$ mm). This test allows for detection of mean differences that are nearly six times smaller than in the first example. Although the same effect sizes exist for both tests, the second test is far more sensitive to the construct of interest.

Another issue is that use of "shirt size" effects often does not correspond to careful thought about the specific problem of interest. Whenever students consult with me on power analysis, I ask what sort of effect size they plan for their design. Most say medium. When questioned, few can justify why they chose that level of effect, other than to say that it sounded like a reasonable compromise between small and large.

Effects Based on Previous Work

Often researchers look to previous work as a guide to estimating effect sizes. Certainly, having some information about effects is better than arbitrarily specifying effects. However, this approach does not address whether or not the effect sizes presented in previous work reflect a meaningful result. Also, it is important to recognize that the effect observed in any single study reflects a sample. This sample effect size may or may not be a good estimate of the population effect size. As discussed earlier, the sample effect is more likely to be an overestimation of the population effect than an underestimation. Published work tends to favor significant results. Studies with larger effects are more likely to produce significant results, so the published literature often overrepresents larger effects.

Effect size estimates derived from a body of literature (e.g., 10 studies examining similar effects found d ranging from .10 to .30) temper concerns about getting a reasonable estimate of the effect, as there is a larger sample of effect sizes used in estimation. In situations like this, overrepresentation of significant (and therefore larger) effects remains an issue, so choosing from the lower end of estimates rather than the upper end (e.g., .10ish rather than .30ish) is a conservative decision that sometimes offsets overrepresentation. Carefully conducted meta-analyses that include unpublished literature reduce sampling concerns and may provide accurate effect size estimates. However, it is important to note that many long-utilized meta-analytic techniques for addressing publication biases now appear to be flawed (Carter, Schöbrodt, Gervais, & Hilgard, 2017). Given these limitations, it may be more reasonable to pick only the well-powered studies found in a meta-analysis as reasonable estimates of the population effect size.

Despite issues with the published literature, examining previous work is a good reality check when used in conjunction with the "meaningful effect" strategy

discussed in the next section. What others found provides some information regarding typical effect sizes for the area of inquiry and helps provide context for interpretation of effect sizes. For example, if effects in your field typically hover around $d = 0.20$, then designing a study to find a similarly small effect is reasonable only if that size of effect is practically meaningful for your research.

Meaningful or Practically Important Effects

The approach that I recommend involves designing for the minimum effect that is practically meaningful. This is sometimes termed the "smallest effect size of interest" (SESOI, Lakens, 2014). The idea of meaningfulness is slippery and often not entirely obvious, particularly for basic areas of research. In thinking about meaningful effects, there are a number of questions to ask but not all the questions may be relevant to every project.

A good beginning question when designing an intervention or similar study is how much improvement the intervention needs to make to justify the cost. For example, McCartney and Rosenthal (2000) showed that an active learning program that produced a small effect of $r = .14$ on improving student learning related to a return of over \$7 for every \$1 spent on the program. In this case, what appears to be a small effect offers considerable gains. Now consider another situation in which a similar program costing 10 times as much produced a similar effect size. For this program, the return is \$0.70 for every \$1 spent. Both programs produce the same effect size, but the first yields greater benefits based on the cost. Continuing the educational program example, another question is how this effect compares to those found for similar programs. If a typical educational program produces only a \$0.25 return on each dollar, then the program with the \$0.70 return would be a bargain.

Unfortunately, for many basic research topics, cost–benefit analyses are not relevant. For this type of work, it is important to become familiar with the published literature in your area. A good approach is to start with a search of the literature to get a sense of typical effects for research in your area, then use those effect sizes to construct an initial effect size estimate. After deriving this initial estimate, it can be useful to translate the standardized effect size to the actual units of interest to get a better understanding of the effect size in practical terms. For example, if examining whether an experimental manipulation reduces anxiety and the literature shows that this form of anxiety reduction produces effects of $d = 0.20$, translating this information into units on the anxiety scale may be easier to interpret than the effect size. If scores on the scale of interest ranged from 1 to 10 with higher scores meaning greater anxiety and the scale had a standard deviation of 2.0, an effect size of $d = 0.20$ would reflect a raw score difference of less than one point. A reasonable question is whether such a small difference is enough to support use of the technique or if the practical value of a difference smaller than 1 point is too small to be of interest.

Effect Sizes for Replication Studies

Replications provide a unique situation for power analysis. There is an empirical estimate of effect size from the previous work. However, given issues with publication bias, this effect is more likely than not an overestimate of the true population effect. For such situations, Simonsohn (2015) suggests setting the smallest effect size of interest to the effect size that the original study had power of 33% to detect. For example, if the original study used a two-group design with 64 participants in each group, the study has power of .33 to detect an effect size of $d = 0.27$ (see Chapter 3 for determination of power under these conditions). Similarly, Lakens (2017a) argues that the original research design reveals the smallest effect size the original researchers cared about detecting. For example, a study designed to detect $d = 0.50$ with $n = 64$ per group and 80% power, will reject the null hypothesis for any sample that yields $d > 0.30$, suggesting that $d = 0.30$ is the smallest effect of interest.

Concluding Comments on Determining Effect Size

There are many ways to estimate effect sizes for power analysis. Unfortunately, there are issues with most approaches. Designing around small, medium, or large effects often reflects arbitrary decision making. Using existing research, even meta-analyses, often results in overly optimistic decisions. Designing around the smallest effect size that is meaningful is challenging and most obvious when engaging in cost–benefit analyses. Despite these challenges, throughout the book I focus on designing to detect meaningful effects.

Post Hoc Power (a.k.a. Observed or Achieved Power)

This book focuses on power analyses as an a priori venture. The value of power analyses is highest before data collection. There are, however, some arguments for providing power analyses for completed work. These approaches, sometimes termed post hoc, achieved, observed, or retrospective power, provide a power estimate based on the effect size observed in the sample and the sample size. Post hoc power analysis therefore tells how much power we had (given our sample size and α) to attain statistical significance if the effect size in our sample is the true population effect. Proponents of post hoc power analysis argue that for nonsignificant results, post hoc power provides useful information about the need for replication, with low power suggesting replication is necessary to draw conclusions about whether or not a Type II error existed (e.g., Onwuegbuzie & Leech, 2004). This perspective suggests that high post hoc power supports the veracity of the failure to reject the null hypothesis conclusions (i.e., provides support for the null hypothesis). That is, we had relatively high post hoc power but still could not reject the null hypothesis.

My view is that post hoc power is not particularly useful. First, power is inversely related to both significance test probabilities and effect sizes. Failing to reject the null hypothesis generally means that post hoc power was low and a larger sample size is needed to obtain statistical significance for the observed effect size (e.g., Lenth, 2001; Nakagawa & Foster, 2004). Thus, post hoc power tells us nothing new. Another flaw in the logic of post hoc power proponents is that power increases as significance test probabilities decrease, meaning that higher levels of post hoc power occur when tests approach significance. Use of power to support null hypotheses therefore employs a procedure wherein results that almost met criteria for rejection of the null hypothesis correspond to more support for null effects (Hoenig & Heisey, 2001). For example, given two analyses with the same sample size, a comparison of two groups that produces $p = .70$ would return an estimate of low power (e.g., .10), whereas a sample producing $p = .08$ (i.e., just missing the criteria for rejecting the null at $\alpha = .05$) would yield substantially more power (e.g., .70). Under this flawed view, the second result would suggest stronger evidence that the null hypothesis was in fact true, as power was higher in this situation.

Most uses of post hoc power estimates likely occur for two reasons. First, programs such as SPSS provide post hoc power estimates (called "observed" power). Second, reviewers sometimes request these values and authors lack the knowledge to argue against such presentation. In short, I do not believe post hoc power estimates provide useful information and along with others (e.g., Maxwell, Kelley, & Rausch, 2008), call for strategies such as CI drawn around effect sizes when focusing on "support" for null hypotheses.

Post hoc power analysis is perhaps best characterized by this quotation "To call in the statistician after the experiment is done may be no more than asking him [or her] to perform a post-mortem examination: he [or she] may be able to say what the experiment died of" (Fisher, 1938, p. 17). That is, if we find that post hoc power is low, all we know is that the observed effect size was too small to be detected with the design we used.

Post hoc Power as a Bias Detection Tool

Since the first edition of this text, some useful approaches utilizing post hoc power emerged. For example, post hoc power informs indices of the credibility of a set of studies in a single paper. Observed power for each of a set of studies can be calculated and then combined to address the overall power of a study to detect effects. For example, if a manuscript reports significant effects supporting predictions across four studies with observed power of .90, .80, .95, and .90, the product of these values inform power for the entire set of studies. Power to detect all of the effects in the same set of studies in this case is .6 (.90*.80*.95*.90), suggesting a somewhat unsurprising set of studies.

There are now numerous examples in the research literature of similar approaches used to question the credibility of existing results. For, example an analysis of Bem's (2011) controversial paper on pre-cognition (the ability to know what will happen in the future), found power of <.001 for a set of 10 studies yielding 19 tests, 14 of which were significant. The author of this work concluded that "it is unlikely that Bem (2011) conducted 10 studies, ran 19 statistical tests of planned hypotheses, and obtained 14 statistically significant results" (Schimmack, 2012, pp. 558–559). For more extensive information on the use of post hoc power in this manner, see Schimmack (2016).

How Much Power?

One remaining question for this chapter is how much power to target. Power of .80 for tests aiming to reject the null hypothesis seems a de facto standard. Many examples in the text design around this 80% value, however, that is not an endorsement of 80% as a meaningful level of power in all circumstances. Whereas 80% is a reasonable level of power for most situations, other considerations need exploration. For example, if the cost of a Type II error were high (e.g., for a treatment that was expensive to develop), then designing for more power would be desirable.

The 80% standard is interesting to investigate. Figure 1.7 shows power for small, medium, and large effect sizes. One thing to note on the graph is that the relationship between power and sample size is roughly linear when moving from power of .20 to .80. However, moving from power of .80 to higher values

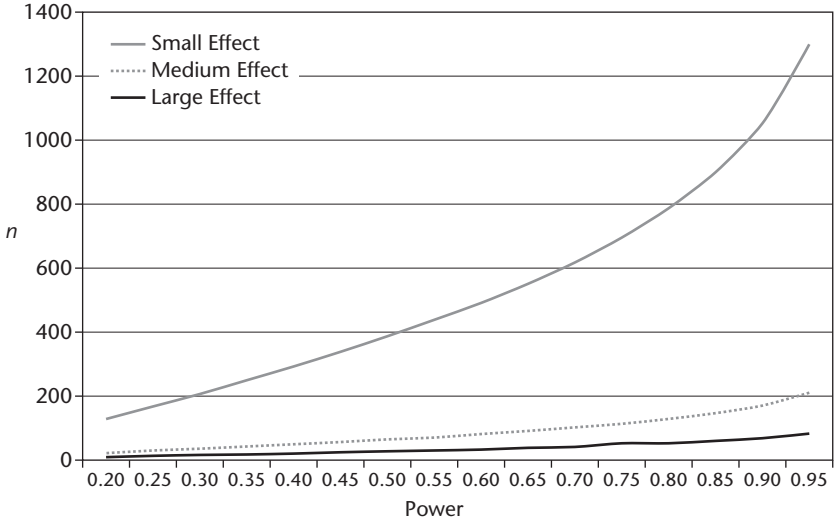

FIGURE 1.7 Sample Size and Power for Small, Medium, and Large Effects.

TABLE 1.3 Percentage and Sample Size Increases for Small, Medium, and Large Effects

Power Increases From	*Small*	*Medium*	*Large*
.50 to .60	26.8% (104)	28.1% (18)	21.4% (6)
.60 to .70	26.0% (128)	24.4% (20)	23.5% (8)
.70 to .80	27.1% (168)	25.5% (26)	23.8% (10)
.80 to .90	33.8% (266)	34.4% (44)	30.8% (16)
.90 to .95	23.5% (248)	22.1% (38)	23.5% (16)

corresponds to a sharp upturn in required sample size (the small effect shows this relationship most clearly). Table 1.3 shows this in terms of the percentage increase in sample size required to increase power.

As shown in Table 1.3, increasing power reflects consistent jumps of about one-quarter of the sample size for moving from .5 to .6, .6 to .7, and .7 to .8. However, moving from .8 to .9 requires an increase of around one-third of the sample size. Moving from .9 to .95 requires another one-quarter increase. This suggests that power of .80 combines the best sample–power balance. However, if you can afford more power, by all means, design for more power. For example, if designing for a large effect and additional participants would not increase costs considerably then a design with power of .90 or even .95 would be advantageous.

Summary

This introductory chapter examined NHST, effect sizes, and basic issues in power analysis. Many of the issues in the present chapter receive extended coverage in later chapters. For example, throughout the text, there are formulae and examples addressing calculation of effect size estimates and noncentrality parameters. Similarly, issues relevant to determining effect size estimates receive coverage throughout the book.

Notes

1. Figure 1.5 was created using the Exploratory Software for Confidence Intervals software program (ESCI). ESCI provides an outstanding visualization tool for exploring distribution shapes (see Chapter 11 for information on obtaining this free software).
2. Figure 1.6 was produced using ESCI.

2

CHI SQUARE AND TESTS FOR PROPORTIONS

Chi square tests for independence (also known as multicategory, contingency) and Goodness of Fit (GoF; also known as one-way classification) address questions regarding distribution of scores into categories. The GoF test involves distributions for categories of a single variable whereas the test for independence examines whether membership in one category relates to membership in another (i.e., an interaction between variables). This chapter provides a detailed example using the test of independence and brief examples for GoF and 3×2 independence tests. The chapter also addresses tests involving comparing a proportion to a hypothesized proportion (one sample test) and tests comparing two independent proportions.

As this is the first chapter focusing on specific techniques, I highlight some general considerations for power analyses, discuss specific tests, and address computer-based approaches. In particular, discussions of noncentrality parameters (NCPs), determination of meaningful effects, and use of R functions for power analysis are relevant to the material in this chapter and many of the designs presented in other chapters.

Necessary Information

The tests covered in this chapter focus on frequencies or proportions. Both forms of χ^2 are concerned with frequencies, those expected based on the null hypotheses and observed values. Estimating power for χ^2 requires construction of the proportions reflecting the alternative distribution and proportions reflecting the null hypothesis. The two tests of proportions (one sample and independent) involve the same type of information, but generally for fewer groups. In either case, the alternative distribution should reflect a population that deviates from the null hypothesis in a manner that would be practically meaningful or interesting.

Factors Affecting Power

Larger deviations between observed and expected values produce larger effect sizes as do larger differences between group proportions. Differences between more extreme proportions (those closer to 0 or 1.0) produce larger effect sizes. For example, the difference between proportions of .8 and 1.0 corresponds to a larger effect size than the difference between .4 and .6. As with any analysis, larger effect sizes, larger samples, and more liberal α levels yield more power.

Key Statistics

This section presents formulae for Chi square tests. Information for other tests of proportions appear in other sections of the chapter. One value that deserves special mention is the NCP. NCPs measure the distance between the distribution from which we are sampling (also known as alternative distribution) and the distribution specified in the null hypothesis. The shape of the alternative distribution reflects the effects we wish to detect and sample size employed so the NCP calculation is driven by the effect size and sample size.

For our purposes, the NCP is an intermediate step to determining power. The NCP value has little immediately interpretable meaning aside from bigger meaning more power. NCPs are available for the *t*, *F*, and χ^2 distributions. *F* and χ^2 use a value called λ (Lambda); *t* uses δ (Delta).

Pearson's Chi Square (χ^2)

The fit of observed frequency data to a model of expected frequencies is commonly assessed with the Pearson's χ^2 statistic. This statistic, shown in Formula 2.1, focuses on two frequency values, observed (f_o) and expected (f_e). The more deviant the observation is from the expectation, the larger the χ^2 value. In short, the approach sums the squared difference between the observed and expected frequencies of each cell over the expected frequency of the cell. The term expected reflects what we would observe if no effect were present. In the context of power analysis, observed takes on a slightly different meaning. Power analysis is a priori, so we do not have data, making the term "observed" a bit of a misnomer. Observed reflects what the population of interest looks like. As discussed in the next section, we are interested in what sort of observed difference would be meaningful. A key question to ask is "how large would the deviation between observed and expected have to be for the result to be interesting or practically important?" When conducting power analysis for Chi square, we establish the proportions (or frequencies) we would observe if there were meaningful differences between groups.

$$\chi^2 = \sqrt{\sum_{i=1}^{m} \frac{(f_{oi} - f_{ei})^2}{f_{ei}}} \tag{2.1}$$

Phi (Φ)

Phi (also known as Cramer's V or *w*) is a common measure of effect size for frequency data. When applied to a 2×2 test of independence, the value is equivalent to Pearson's correlation coefficient (*r*). Φ focuses on observed and expected proportions rather than frequencies. In comparing the χ^2 and Φ formulae, we see that Φ is simply χ^2 expressed proportionally rather than based on frequencies (technically, the square root of those values). That is, it is the significance test statistic (χ^2) with all of the elements relevant to sample size removed. The second version of the formula (Formula 2.3) demonstrates this concept more directly.

$$\Phi = \sqrt{\sum_{i=1}^{m} \frac{(p_{oi} - p_{ei})^2}{p_{ei}}} \tag{2.2}$$

$$\Phi = \sqrt{\frac{\chi^2}{n}} \tag{2.3}$$

Lambda (λ)

The NCP for χ^2 is λ. This is a function of Φ and sample size. For power analyses, λ reflects the χ^2 we would obtain if a sample result reflected the expected values exactly.

$$\lambda = n\Phi^2 \tag{2.4}$$

Example 2.1: 2 × 2 Test of Independence

In this example, I present power analysis for a replication study addressing effects of a prospective tenant's reference to HIV on responses to inquiries about the availability of rooms or apartments for rent (Page, 1999). In the original study, callers inquired about availability of rentals and either mentioned or did not mention undergoing treatment for HIV. The outcome measure was whether the rental was reported as still available or not. The study found that 40% of applicants who mentioned HIV were told the rental was available compared to 76% of applicants who did not mention HIV.

In replicating this study, we address how large a sample size is necessary to produce power of .80. The first step in this process is to determine what size of effect our design will employ. There are two approaches to determining effect size detailed next.

Effect Sizes Based on Previous Research

Determining effect sizes based on previous research asks the question "how large a sample is needed to detect a population effect size equivalent to the effect

size in the previous study?" This is a common approach to power analysis but not one that I can endorse without considerable qualification. Estimates based on a single sample may not accurately represent the population effect size. Published research tends to favor statistically significant findings (Rosenthal, 1979). Significant findings tend to have larger effect sizes, so reliance on previously published work often overestimates the population effect size.

When designing a study based on effect sizes from similar studies, it is important to recognize that what someone else found may not be relevant to whether the size of effect you want to detect is meaningful or practically valuable. Of course, this does not mean that the approach is useless. What others found may represent a meaningfully sized effect or the smallest effect that you are interested in detecting. Do not accept mindlessly what other researchers found as a reflection of the size of effect you want to detect. Instead, focus on meaningful effects when possible.

Effect Sizes Based on Detecting Meaningful Effects

The approach that determines effect size based on detecting meaningful effects (i.e., the smallest effect of interest) begins with the question "how large a sample is necessary to detect the smallest effect we would term meaningful?" This approach requires important decisions about what we consider a meaningful difference. This is not always an easy question to answer. Addressing this question adequately involves serious thought and can promote large sample requirements, especially if the researcher believes small differences are important to detect.

It is difficult to provide systematic guidance for determining meaningful effects. Every research study is different with unique concerns and outcomes. Questions about what is or what is not a meaningful result must be answered in the context of each project. I provide rationale for the choices made for the determination of meaningful results and general strategies for answering questions of meaningfulness but ultimately only the researcher can answer this question about their study.

In thinking about what is a meaningful result, it is sometimes useful to first think about data in terms of raw scores (e.g., frequencies, means) rather than as standardized effect sizes. This differs from classic approaches to power analysis that focus first on effect size (e.g., Cohen, 1988; Kraemer & Thiemann, 1987). Thinking about standardized effect sizes often removes the context of effects, making it difficult to determine what is meaningful. In discussing statistical reporting of effects, Wilkinson and the Task Force on Statistical Inference noted "if the units of measurement are meaningful on a practical level … then we usually prefer an unstandardized measure to a standardized measure" (1999, p. 599). Unstandardized measures are generally easier to think about and understand than are standardized measures. This does not mean that standardized effect sizes are not useful, only that it can be easier to begin by examining raw statistics.

Example of Determining Effect Size for Analysis

Going back to the rental study, a good place to start is by examining differences in rental availability between HIV positive applicants and a control group. A good beginning question is "how much discrimination would be meaningful?" Consideration of the level of discrimination found in the earlier study is a good starting point. Page (1999) found a 36% difference in rental availability between HIV positive applicants and the control group (40% vs. 76%). It would be reasonable to term this level of bias "meaningful." A large difference in availability reflects clear discrimination. However, if we designed for power of .80 to detect this much bias (36% difference), we would have considerably less power to detect smaller differences in bias. Unless a 36% difference is the smallest meaningful difference we were interested in detecting, designing around this value would not be a good approach.

Given the issues discussed above, an important question is "how large a difference is meaningful?" Certainly, any amount of discrimination is troubling, so we could design a study that allowed for detection of very small differences (e.g., 5%). However, this small a difference does not seem particularly large when compared to the previous finding of a 36%. As a compromise, I am going to define a 20% difference as the smallest difference I am interested in having power to detect.

For Chi square, the effect size is a function of the proportions. Differences between more extreme proportions produce larger effect sizes than differences between proportions closer to .50. For that reason, it is important to establish proportions based on reasonable estimates for each group rather than simply choosing two proportions that create specific levels of difference. A good starting point is the information from the study about the control group. In the previous study, 76% of the rentals in the control group were available. This value serves as a reasonable baseline, as there does not appear to be any reason to believe availability for this group would change substantially. Table 2.1 shows how the 20% difference is applied to construct proportions corresponding to these values. Recall that p_o (proportion observed) establishes the meaningful differences whereas p_e (proportion expected) describes the null hypothesis.

TABLE 2.1 Proportions Reflecting a Meaningful Difference (Null in Parentheses)

	HIV/AIDS Mentioned (Treatment) p_o (p_e)	*% of Condition*	*No HIV/AIDS Mentioned (Control)* p_o (p_e)	*% of Condition*
Rental Available	.28 (.33)	56	.38 (.33)	76
Rental Not Available	.22 (.17)	44	.12 (.17)	24

A few notes on the values in Table 2.1 are necessary. The proportions reflect the overall proportions of the sample with half of the participants assigned to the control group and half assigned to the treatment group. The original study found 76% of rentals available to those who did not mention HIV. These cells are set at .38 and .12. The value of .38 is the proportion of the total that corresponds to 76% of the control group (recall the control group is only half of the entire sample). The value of .12 is the proportion corresponding to the 24% of the control group for whom the rental was not available. Treatment group proportions are set as deviations from those values. Thus, a 20% difference here reflects a .10 difference in proportions across columns. The 20% difference reflects a difference in terms of how many people in each condition were told the rental was available. The proportional values are out of the total sample, with an equal division of participants between treatment and control groups. The expected proportions reflect the average of the two proportions in the row, or more simply what the proportions would look like if there were no differences between treatment and control. These proportions are especially important as these are the values used for p_e in calculations. Based on the proportions in Table 2.1, Φ is calculated using Formula 2.2.

$$\begin{aligned}\Phi &= \sqrt{\sum_{i=1}^{m} \frac{(p_{oi} - p_{ei})^2}{p_{ei}}} \\ &= \sqrt{\frac{(0.28-0.33)^2}{0.33} + \frac{(0.38-0.33)^2}{0.33} + \frac{(0.22-0.17)^2}{0.17} + \frac{(0.12-0.17)^2}{0.17}} \\ &= 0.2111\end{aligned}$$

The effect size is then used to calculate the NCP. For this calculation, we need to choose a sample size as a starting point. The calculations (using Formula 2.4) show that with $n = 100$, λ is equal to 4.41.

$$\lambda = n\Phi^2 = 100(0.2111)^2 = 4.46$$

As a short aside, it appears that thinking about these differences in terms of percentages makes a bit more sense than beginning with an effect size. For example, a study designed to detect a minimum of a 20% difference in availability (56% vs. 76% for HIV and control respectively) would produce $\Phi = .21$. In this case, the percentage result is more intuitive to most than the effect size. Please note that my focus on proportions instead of effect sizes is not a criticism of the utility of effect size measures. Effect sizes are indispensable measures but are not always an easily interpretable starting point for determining meaningful effects for power analysis. One note of caution is that the percentage difference expressed here (20%) may correspond to different effect sizes. Percentage difference and effect size do give different information. For example, if the proportions of interest were .80 and 1.0, respectively, the effect size would be considerably larger ($\Phi = .33$).

Using the Noncentrality Parameter to Calculate Power

Given an effect size of $\Phi = .21$ we can compute power. There are two questions addressed in this section. First, how much power does $n = 100$ yield? Second, what sample size produces Power = .80?

For the first question, we need λ (4.46) and a critical value for χ^2 with our desired α. Using $\alpha = .05$ corresponds to a critical value of $\chi^2 = 3.84$ for $df = 1$. The degrees of freedom are Rows-1 times Columns-1 for this test. Since power calculations for Chi square distributions require use of a noncentral distribution, this calculation requires a computer protocol (see Chapter 1's discussion of this issue).

Using R, the following command computes power:

```
1-pchisq(Chi-Table, df, λ)
```

The value called Chi-Table reflects the critical value for χ^2 (3.84). The $df = 1$ and Lambda = 4.46 so the command looks like this:

```
1-pchisq(3.84, 1.4, 4.46)
```

The pchisq function in R performs calculations based on noncentral distributions; the value it gives is the area below λ. Because R gives the area below λ, the command takes 1 minus the result, as this corresponds to the area above λ that gives power. Using this calculation R yields Power = .56.

Approximating Power Calculations

Although it is not possible to calculate correct power estimates for χ^2 by hand, good approximations are possible. I include this section as my experience teaching about power suggests that approximation, albeit not always accurate, facilitates theoretical understanding of power. The approach demonstrated here works only when $df = 1$. This approach does not work particularly well for small samples or small effect sizes. As discussed in Chapter 1, with large samples, central and noncentral distributions look similar. The approximation procedure relies on central distributions.

Calculating approximate power directly from λ is relatively simple when $df = 1$. With $df = 1$ the χ^2 distribution is simply the normal distribution squared (z^2). Converting λ and the critical value of χ^2 to z involves taking the square root of each value. Then take these values to Formula 2.5, that yields a z statistic that can be used to approximate power.

$$z_{power} = z_{critical} - z_{\lambda} \tag{2.5}$$

For this example, the critical value for χ^2 with $df = 1$ and $\alpha = .05$ is 3.84. Converted to z, we get $z_{critical} = 1.96$. λ converts to $z_{\lambda =} 2.10$. Inserting these values

into Formula 2.5 yields $z_{power} = -0.14$. This value reflects the point on the alternative (or true) z distribution that corresponds to the critical value of the null distribution. The area below z_{power} reflects samples that do not allow for rejection of the null hypothesis. This corresponds to β or Type II error. The area above z_{power} reflects samples that allow for rejection of the null. This is $1-\beta$ or power.

$$z_{power} = 1.96 - 2.10 = -0.14$$

Taking z_{power} to a normal distribution table shows that 56% of the distribution falls at or above $z = -0.14$. Figure 2.1 shows these calculated values. In this figure, power is the area of the alternative distribution that falls to the right $z_{critical}$. This area is a little above 50% (56% for this example). The value for z_λ represents the center of the alternative distribution (Formula 2.6).

The approximation technique also allows for the determination of a sample size corresponding to a particular level of power. For example, if we want to find Power = .80 (or some other value), we can work backward and rearrange the formulae. First, find the z-value above which you find 80% of the distribution. This corresponds to $z = -0.85$. Plug that value into Formula 2.6 with the critical value for z_λ and solve for z_λ. Square z_λ to get λ. Finally, take λ to Formula 2.7 and solve for n. In this case, for Power = .80, a sample of 179 is necessary. To facilitate assignment of equal numbers of participants to the two groups round up to $n = 180$.

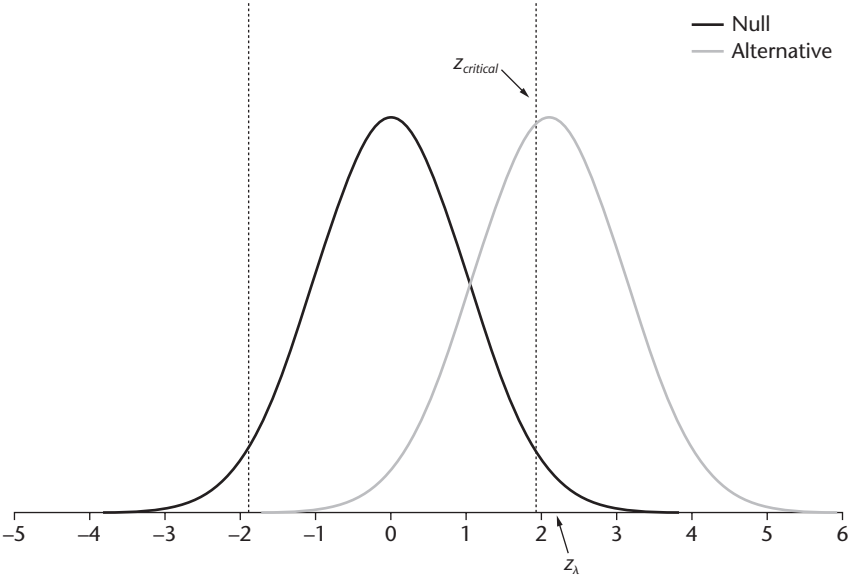

FIGURE 2.1 Graph of Power Area Using Normal Distribution Approximation.

$$z_\lambda = z_{critical} - z_{power}$$

$$z_\lambda = 1.96 - (-0.85) = 2.81$$

$$(z_\lambda)^2 = (2.81)^2 = 7.90$$

$$n = \frac{\lambda}{\Phi^2} \tag{2.7}$$

$$n = \frac{7.90}{0.21^2} = 179$$

Example 2.2: 2 × 2 Chi Square Test for Independence Using R

The calculations detailed in this chapter are effective but impractical for complex designs. The R functions demonstrated in this book (and found in the pwr2ppl package) generally require a single line of code to run analyses. Most R functions return power for a specific sample size. This often necessitates running analyses several times, and trying out new sample size values to hone in on the solution that provides the desired level of power.

The Chi2x2 function presented in Table 2.2 calculates power for a sample of $n = 100$ using the proportions in Table 2.1 and $\alpha = .05$. The format of the function is:

Chi2x2(r1c1, r1c2, r2c1, r2c2, n, alpha)

For the function, r1c1, r1c2, r2c1, and r2c2 refer to the proportion of overall scores in each row by column combination. For example, r1c1 is .28, reflecting the first row and column in Table 2.1. Sample size is defined by n. Alpha defaults to .05 if no value is entered.

Table 2.2 shows the output for an initial analysis based on 100 participants, yielding power of .56 (just as calculated earlier).

To determine the sample size needed to achieve a desired level of power (.80) with the specified effect size, change the value for n in the command compute $n = 100$ to another value. The basic approach to using the R commands is to run the analysis with a specific sample size and then to run again with a new sample until achieving the desired level of power. Table 2.3 shows the code and output for two analyses. For the first analysis, power was .56 with $n = 100$, so we try a larger sample size and then adjust the sample size as necessary until the reported power reaches the desired level. With $n = 180$, power is .808.

TABLE 2.2 R Code and Output for χ^2 Independence Test ($n = 100$)

```
Chi2x2(r1c1=.28, r1c2=.22, r2c1=.38, r2c2=.12, n=100)
## [1] "Power for n of 100 = 0.56"
```

TABLE 2.3 R Code and Output for χ^2 Independence Test ($n = 180$)

```
Chi2x2(r1c1=.28, r1c2=.22, r2c1=.38, r2c2=.12, n=180)
## [1] "Power for n of 180 = 0.808"
```

Example 2.3: Other χ^2 Tests

Brief examples of GoF and independence tests with more than two categories appear in Table 2.4.

Goodness of Fit

The ChiGOF function in Table 2.4 addresses power for the GoF test. The format of the function is:

ChiGOF(po1, po2, po3, po4, po5, po6, groups, n)

The po values reflect proportions in each group. Two is the minimum and six is the maximum. Proportions must add to 1.0. The values groups and *n* define number of groups and sample size, respectively.

In this example, there is a single variable with four categories. The function tests a null hypothesis of equal proportions across groups. The analysis compares a population where meaningful differences between categories reflect proportions of .25, .20, .20, and .35 to a null distribution that represents equal proportions across the groups.

Test of Independence with More Than Two Categories

For this example, the only change from the 2 × 2 is the addition of a third category for the second variable. The function now requires specification of proportions for r1c3 (row 1, column 3) and r2c3 (row 2, column 3).

Example 2.4: General Effect Size-Based Approaches Using R

An additional tool in the pwr2ppl package is a function called ChiES. The format of this function is:

ChiES(phi, df, nlow, nhigh, by, alpha)

TABLE 2.4 R Code and Output for Goodness of Fit and 2×3 Independence Test

Goodness of Fit	*2 × 3 Independence*
`ChiGOF(po1=.25, po2=.20, po3=.20, po4=.35, groups=4, n=100)` `## [1] "Power for n of 100 = 0.518"`	`Chi2x3(r1c1=.25, r1c2=.25, r1c3=.10, r2c1=.10, r2c2=.25, r2c3=.05, n=200)` `## [1] "Power for n of 200 = 0.747"`

TABLE 2.5 R Code and Output for General Effect Size Approach to Power for Chi Square

```
ChiES(phi=.3, df=1, nlow=10, nhigh=200, by=10, alpha=.01)
## [1] "Power for n of 10 = 0.0521"
## [1] "Power for n of 20 = 0.1086"
## [1] "Power for n of 30 = 0.1755"
## [1] "Power for n of 40 = 0.2487"
## [1] "Power for n of 50 = 0.3247"
## [1] "Power for n of 60 = 0.4005"
## [1] "Power for n of 70 = 0.4737"
## [1] "Power for n of 80 = 0.5428"
## [1] "Power for n of 90 = 0.6065"
## [1] "Power for n of 100 = 0.6643"
## [1] "Power for n of 110 = 0.7159"
## [1] "Power for n of 120 = 0.7613"
## [1] "Power for n of 130 = 0.8009"
## [1] "Power for n of 140 = 0.8349"
## [1] "Power for n of 150 = 0.864"
## [1] "Power for n of 160 = 0.8886"
## [1] "Power for n of 170 = 0.9092"
## [1] "Power for n of 180 = 0.9263"
## [1] "Power for n of 190 = 0.9405"
## [1] "Power for n of 200 = 0.9522"
```

The value phi reflects the effect size (Φ). Degrees of freedom are defined by *df*. nlow is the starting sample size and nhigh the ending size. The value by defines the incremental increase from nlow to nhigh with a default of 1 unit. Alpha defaults to .05, as with the other functions alpha can be set to other values (e.g., alpha = .01). See Table 2.5 for an example.

Tests for Single Samples and Independent Proportions

In addition to Chi-square approaches to testing for differences in proportions, several other techniques compare proportions. This section presents power analysis for tests comparing a sample to a hypothesized value for the population (single sample test) and comparing proportions between two independent populations. Calculations for both approaches are similar, so I present general formulae first then address application to each design.

Formulae for Differences in Proportion Tests

The effect size for tests involving differences in proportions is termed *h*. The value *h* is like *d* and can be thought of (at least intuitively) in the same manner (i.e., units of standard deviation). The calculation of *h* involves what is

commonly termed the arcsine transformation. This transformation deserves some special mention. The name arcsine transformation is imprecise, sometimes leading to incorrect use of the transformation, likely because of confusion over the proper calculation. For example, MS Excel provides an arcsine function that does not correspond to the transformation in Formula 2.8. MS Excel correctly calculates the arcsin portion of the formula but do not include the 2 or square root of *p* parts.

Formula 2.8 notes the arcsine transformed proportion as p'. Others (e.g., Cohen, 1992) note this value as Φ. I avoid this notation, as Φ is also the effect size for Chi square. The calculation is applied to two proportions of interest, either the alternative and null proportions or two independent proportions.

$$p' = 2\arcsin\sqrt{p} \tag{2.8}$$

After converting both proportions and the effect size, calculate *h* by subtracting one transformed proportion from the other (Formula 2.9).

$$h = p_1' - p_2' \tag{2.9}$$

There is not an NCP for this test as the difference between proportions is tested against the normal distribution and such tests use a central distribution. Formulae 2.10 and 2.11 use z_λ to note the value that is analogous to the NCP in the previous parts of the chapter. The choice between Formulae 2.10 and 2.11 depends on the test. For a one sample test, use Formula 2.10. For a test of independent samples, use Formula 2.11. For unequal samples sizes among independent groups, use the harmonic sample size defined by Formula 2.12. Once z_{power} is calculated, we apply Formula 2.5. Since this test uses the normal distribution (rather than chi square as seen in the previous examples), accurate hand calculation of power is possible.

$$z_\lambda = h\sqrt{n} \tag{2.10}$$

$$z_\lambda = h\sqrt{n/2} \tag{2.11}$$

$$n_{harmonic} = \frac{2n_1n_2}{n_1 + n_2} \tag{2.12}$$

Example 2.5: Single Sample Comparison

As a graduate student, I worked coding studies for a large-scale meta-analysis involving drug abuse treatment programs. One of the outcomes of interest was the proportion of program participants who remained abstinent for a certain period (e.g., 6 months). This was one of several possible outcomes, often with so many categories that the data were not appropriate for χ^2. For

example, categories might include abstinent 1 year, mostly abstinent for a year (one or two slip-ups), abstinent 6 months, abstinent 3 months, abstinent 1 month, dropped out of treatment, and not abstinent. Two designs for examining effectiveness of programs involved comparing a sample from the program to a program goal for abstinence or comparing two samples of participants who received different forms of treatment. For example, a program might compare its abstinent rates to a benchmark for abstinence (single sample comparisons) or abstinence rates of a sample drawn for a comparison program (independent samples comparisons).

Imagine that a program would qualify for extended funding if it demonstrated substantial improvement over a 42% success rate for routine treatments as reported in previous meta-analysis of similar programs.[1] The program believes that, at minimum, they had a 60% success rate. How large a sample of program participants is necessary to provide power of .90 to detect this difference? This approach involves a one sample proportion test as there is a sample proportion compared against a hypothesized value.

Calculating the effect size involves transforming both proportions (.60 and .42), then taking the difference between the two transformed proportions. Next, the effect size can be used to calculate z_λ, which is then used to obtain power. The examples that follow use an arbitrary sample size of 20 and a one-tailed $\alpha = .05$ test.

$$
\begin{aligned}
p_1' &= 2\arcsin\sqrt{p} = 2\arcsin\sqrt{0.60} = 1.77 \\
p_2' &= 2\arcsin\sqrt{p} = 2\arcsin\sqrt{0.42} = 1.41 \\
h &= p_1' - p_2' = 1.77 - 1.41 = 0.36 \\
z_\lambda &= h\sqrt{n} = 0.36\sqrt{20} = 1.61 \\
z_{power} &= 1.645 - 1.61 = 0.035
\end{aligned}
$$

The area above z_{power} of 0.035 is .49, indicating that we have Power = .49, well below our desired level.

Table 2.6 demonstrates use of the prop1 function and output. The format of the function is:

```
prop1(p1, p0, nlow, nhigh, tails, by)
```

The p values represent the alternative (p. 1) and null (p. 0) proportions. Tails defines a one vs. two-tailed approach with two tails the default. Other aspects of the function such as nlow, nhigh, by, and alpha are largely the same as previous examples.

The function returns power for a range of sample sizes, showing that a sample of around 70 participants yields the desired power.

TABLE 2.6 R Code and Output for One Sample Proportion Tests

```
prop1(p1=.60, p0=.42, nlow=20, nhigh=100, tails=1, by=10)
## [1] "Power for n of  20 = 0.4897"
## [1] "Power for n of  30 = 0.6324"
## [1] "Power for n of  40 = 0.7405"
## [1] "Power for n of  50 = 0.82"
## [1] "Power for n of  60 = 0.8769"
## [1] "Power for n of  70 = 0.9169"
## [1] "Power for n of  80 = 0.9445"
## [1] "Power for n of  90 = 0.9633"
## [1] "Power for n of 100 = 0.9759"
```

Example 2.6: Independent Proportions Comparison

Imagine that instead of comparing against a benchmark, another program wanted to compare the effectiveness of their program to a value-added version of the program that included new therapeutic components. The current program reports a success rate of 55%. Given the increased costs associated with the new program, they would consider the value-added program worthwhile if it were to, at minimum, increase success to 62%. By worthwhile this would mean that the level of improvement would be enough to justify a full-scale change in the program. As the new program is experimental, only a proportion of program participants (20%) are scheduled to participate in the value-added program.

Comparisons of independent samples for proportions proceed in the same fashion as the single sample test with some additional considerations. The effect size statistic, h, is calculated the same way. The calculation of z differs slightly and groups have different sample sizes. With unequal sample sizes, the harmonic sample size replaces n in the calculation of z. The following example uses sample of 200 participants (20% assigned to the new treatment).

$$p_1' = 2\arcsin\sqrt{p} = 2\arcsin\sqrt{0.62} = 1.813$$

$$p_2' = 2\arcsin\sqrt{p} = 2\arcsin\sqrt{0.55} = 1.671$$

$$h = p_1' - p_2' = 1.813 - 1.671 = 0.142$$

$$n_{harmonic} = \frac{2n_1n_2}{n_1 + n_2} = \frac{2*40*160}{40+160} = 64$$

$$z_\lambda = h\sqrt{\frac{n}{2}} = 0.142\sqrt{\frac{64}{2}} = 0.80$$

$$z_{power} = 1.96 - 0.80 = 1.16$$

Taking the value of 1.16 to a normal distribution table, shows power is .12 for a sample of 200 program participants (.12 is the area at or above $z = 1.16$).

TABLE 2.7 R Code and Output for Independent Samples Proportion Tests

```
propind(p1=.62, p2=.55, nlow=200, nhigh=2500, by=100, nratio=.2)
## [1] "Power for sample sizes of  40  160 = 0.1239"
## [1] "Power for sample sizes of  60  240 = 0.1648"
## [1] "Power for sample sizes of  80  320 = 0.2054"
## [1] "Power for sample sizes of 100  400 = 0.2457"
## [1] "Power for sample sizes of 120  480 = 0.2855"
## [1] "Power for sample sizes of 140  560 = 0.3245"
## [1] "Power for sample sizes of 160  640 = 0.3627"
## [1] "Power for sample sizes of 180  720 = 0.3999"
## [1] "Power for sample sizes of 200  800 = 0.4359"
## [1] "Power for sample sizes of 220  880 = 0.4707"
## [1] "Power for sample sizes of 240  960 = 0.5041"
## [1] "Power for sample sizes of 260 1040 = 0.5362"
## [1] "Power for sample sizes of 280 1120 = 0.5668"
## [1] "Power for sample sizes of 300 1200 = 0.596"
## [1] "Power for sample sizes of 320 1280 = 0.6237"
## [1] "Power for sample sizes of 340 1360 = 0.65"
## [1] "Power for sample sizes of 360 1440 = 0.6748"
## [1] "Power for sample sizes of 380 1520 = 0.6982"
## [1] "Power for sample sizes of 400 1600 = 0.7203"
## [1] "Power for sample sizes of 420 1680 = 0.741"
## [1] "Power for sample sizes of 440 1760 = 0.7605"
## [1] "Power for sample sizes of 460 1840 = 0.7787"
## [1] "Power for sample sizes of 480 1920 = 0.7958"
## [1] "Power for sample sizes of 500 2000 = 0.8117"
```

Table 2.7 demonstrates use of the propend function. The format of the function is:

propind(p1, p2, nlow, nhigh, by, nratio, alpha)

The values p1 and p2 define group proportions. nratio establishes the balance of participants per group with equal sample sizes the default (nratio = .5). Other aspects of the function such as nlow, nhigh, by, and alpha are the same as previous examples.

Based on the results in Table 2.7, detecting this effect with Power = .80 requires a sample of between 2400 and 2500 participants. Although this is a large sample, the result should not be surprising as the *h* statistic is very small.

Additional Issues

The Chi square test for independence is for nominal categories. If categories are ordinal (e.g., class rank) other measures such as gamma may be more

appropriate. Siegel and Castellan (1988) provide a good discussion of power for nonparametric tests, but certainly more work is necessary in this area.

Violations of χ^2 assumptions occur when cells have very small expected frequencies (e.g., less than 5). For tests of independence, small expected frequencies are a product of low observed frequencies across a category level. For example, a 2 (Condition: Treatment vs. Control $\times$ 3 (Response: Yes vs. Maybe vs. No) design might produce a very small proportion of "No" responses across both conditions, leading the expected frequencies for "no" cell to be very low. One solution is to collapse categories to address this problem (e.g., combine no and maybe responses into a single category). Collapsing this way will turn a 2×3 design into a 2×2. The simpler 2×2 design often will have more power. However, if you expect a problem of this nature, it is a good practice to evaluate power for both the 2×2 and 2×3 designs, using whichever yields a larger sample size requirement (provided that assumptions are met in both cases).

Summary

This chapter addressed power for Chi square tests of independence and GoF, tests involving a single proportion compared to a hypothesized value, and comparisons of two independent proportions. For each design, power analysis involves specifying the null and alternative distributions. The null distribution reflects the proportions specified in the null hypothesis (e.g., equal proportions across groups). The alternative distribution establishes the proportional differences we wish to test relative to the null distribution (e.g., a 10% difference between two groups). Ideally, proportions reflect the smallest difference the researcher defines as meaningful. Relevant to these analyses, this chapter includes formulae and R functions for calculating power using the pwr2ppl package. The chapter also includes a discussion of determining effect sizes for design through consideration of the size of a meaningful difference as well as examination of previous research findings.

Note

1. This success rate comes from Prendergast, Podus, Chang, & Urada (2002) but represents a considerable simplification of the factors contributing to success rate.

3

INDEPENDENT SAMPLES AND PAIRED *t*-TESTS

This chapter focuses on power for designs traditionally addressed using *t*-tests (either independent or paired). These procedures often examine treatment-control group comparisons and pre-post designs. I present power analyses for independent and paired designs with R functions for primary analyses and for tests addressing violation of homogeneity of variances assumptions and unequal sample sizes.

Necessary Information

For designs using independent sample *t*-tests, the initial step is determining means (μs) representing meaningful differences between groups and making a reasonable estimate of standard deviation (σ) for both groups. For paired *t*-test designs, means representing meaningful differences, standard deviation, and the expected correlation (ρ) between dependent measures are required. For the paired *t*-test, the standard deviation of the difference may be substituted for σ and ρ; however, it is usually easier to focus on standard deviations and the correlation.

For both tests, you may also start with an estimate of the effect size. The effect size estimate most commonly used for two-group designs is Cohen's *d*. Technically, power analysis involves a population effect size, which is usually noted as δ or Δ (lower- and upper-case delta, respectively). This text uses *d* to designate the population effect size. This is because δ represents the noncentrality parameter (NCP) (discussed in the next section), and it is confusing to use the same symbol for two different values.

Estimation of the standard deviation deserves special mention. Estimating the population standard deviation (σ) is sometimes tricky. One approach is to

pretest. In general, a pretest with even as few as 20 participants helps to establish a reasonable estimate of variability for the dependent measure.[1] Of course, there are other benefits to pretesting such as establishing whether manipulations actually work and if measures make sense to participants. Another strategy is estimating the standard deviation of the dependent measure from previous uses of the measure. This can be accurate when dealing with established measures. However, this approach requires close attention to the study populations used. For example, a standard deviation based on college students might not represent an accurate estimate of the standard deviation for office workers.

Finally, when estimating standard deviations, it is important to consider potential differences between treatment and control group variability. For example, if group assignment is nonrandom (e.g., samples from existing groups), groups may differ in terms of standard deviations. Unequal variances across groups influence power and complicate estimates of necessary sample sizes. Anticipating these issues in the design stage produces more accurate power analyses.

Factors Affecting Power

For between subjects designs addressed using the independent samples *t*-test, the mean difference and standard deviations influence power. The mean difference and the pooled standard deviation comprise the effect size. For within subjects designs addressed using the paired *t*-test, the correlation between the two administrations of the dependent measure also affects the effect size, with correlations of greater than .50 increasing power and correlations of less than .50 decreasing power.

As noted in Chapter 1, sample size, desired alpha error, and the directionality of the test (i.e., one or two-tailed) affect power. Use of one-tailed (directional) or two-tailed (nondirectional) tests deserves some special mention. A one-tailed approach yields a more powerful test when outcomes are in the predicted direction. For example, a directional test where the expectation is that the treatment group outperforms the control group has more power than a nondirectional test if the actual study results find the treatment group outperformed the control. If it were the case that the control outperformed the treatment group, the directional test would have no power to detect this effect but the nondirectional test retains some power.

The choice of a one- or two-tailed test is an a priori decision. You cannot peek at the data to see which test allows the best conclusion. Often, the decision between one- or two-tailed tests hinges on one key question: "Do I care if the results are opposite what I expect?" In many situations, researchers want to be able to discuss findings in either direction. I rarely see the use of one-tailed tests in the literature in Psychology, likely because most statistical software defaults to a two-tailed probability for most tests. Recent criticisms

of the evidentiary value of probability values just below .05 (two-tailed; Benjamin et al., 2017) suggest increased pressures against using more liberal one-tailed tests.

Key Statistics

This section presents statistical formulae for the analyses that follow. As discussed later in the chapter, R functions perform most of these calculations. When possible, I provide general formulae relevant to both independent and paired samples tests.

Independent Samples t

The *t*-statistic (Formula 3.1) reflects the difference between the sample means minus the hypothesized difference over the standard error of the differences (Formula 3.2). The hypothesized difference between means is set at zero (this is the default in most statistical packages). Test of nonnil hypothesis (e.g., Thompson, 1998), are not discussed in this chapter, but are easily accommodated by simply changing the hypothesized difference (represented in Formula 3.1 as 0) to the nonnil value of interest. The nonnil approach is especially useful for analyses that seek a particular effect size/statistical significance combination (e.g., conclude with a reasonable degree of certainty that our samples do not come from a population where the true difference between means is 2.5).

The standard error of the differences focuses on the standard error of each group and the correlation between measures. For independent samples, $r = .00$ in Formula 3.2

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_{\bar{x}_1 - \bar{x}_2}} \text{ (for samples)} \tag{3.1}$$

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2_{\bar{x}_1} + s^2_{\bar{x}_2} - 2rs_{\bar{x}_1} s_{\bar{x}_2}} \text{ (for samples)} \tag{3.2}$$

Formulae 3.1 and 3.2 use sample notation for calculations of *t* and the standard error of the differences between means. The formulae that follow reflect population values. The use of population values serves as a reminder that the values used for calculations are not based on data collection. Power analysis focuses on expected or meaningful values for a population determined before data collection. The effect size, Cohen's *d*, as noted in Formula 3.3 is the mean difference over the pooled standard deviation (Formula 3.4).[2] This value is also known as the standardized mean difference.

$$d = \frac{\mu_1 - \mu_2}{\sigma_p} \tag{3.3}$$

$$\sigma_p = \sqrt{\frac{\sigma_1^2(n_1 - 1) + \sigma_2^2(n_2 - 1)}{n_1 + n_2 - 2}} \tag{3.4}$$

Formula 3.5 details calculation of the NCP. The value noted as n_j reflects the number of people per group. The section on unequal variances and sample sizes presents formulae addressing designs where group sizes differ.

$$\delta = d\sqrt{\frac{n_j}{2}} \tag{3.5}$$

Paired Samples t

The paired samples *t* approach uses the same values as the independent samples approach with the exception of the effect size and NCP. For the effect size, shown in Formula 3.6, the denominator is the standard deviation of the differences (sometimes written as σ_D). Note that the standard deviation of the differences ($\sigma_{x_1 - x_2}$) is not the same value as the standard error of the differences ($s_{x_1 - x_2}$), one has xs and the other has means as the subscript (*x*-bars) (Formula 3.7).[3]

$$d = \frac{\overline{x}_1 - \overline{x}_2}{s_p} \tag{3.6}$$

$$\sigma_{x_1 - x_2} = \sigma_p\sqrt{2(1 - \rho)} \tag{3.7}$$

Formula 3.8 is the NCP for the paired *t*-test.

$$\delta = d\sqrt{n} \tag{3.8}$$

Approximating Power

Following calculation of the NCP, that value, along with the critical value of *t*, goes in Formula 3.9 to produce a value I term t_{power}. The t_{power} value allows for calculation of approximate power. As noted in Chapter 1, estimation is a good way to get a conceptual understanding of power. However, estimation does not always yield accurate power values. The approach detailed here is an excellent teaching tool, but I recommend the computer-based approaches detailed in the sections that follow for formal analyses.

The value called t_{power}, reflects the area on the alternative distribution above which we can reject the null hypothesis. This means that negative values of t_{power} produce more power than positive ones.

$$t_{power} = t_{critical} - \delta \tag{3.9}$$

Exact Power

Obtaining exact power calculations requires a computer protocol. The following R command performs this calculation.

$$1\text{-pt}(t_{critical}, df, \delta)$$

Other sections in this chapter present formulae for additional topics, such as unequal variances and solving for desired sample size as well as R code for completing all of the calculations presented in the chapter.

A Note about Effect Size for Two-Group Comparisons

Many resources on power analysis begin with an estimate of effect size (e.g., Cohen, 1988). The application of these procedures in conjunction with the publication of shortcut guides (e.g., Cohen, 1992) sometimes focuses researchers on thinking in terms of small, medium, or large effects (corresponding to $d = 0.20$, 0.50, and 0.80, respectively) and addressing power based on these estimates. It is not always useful to focus on effect size at the outset of the research design stage. Although others criticize this "shirt size" approach (e.g., Lenth, 2001) for theoretical reasons, my objection is practical. It is often easier for researchers to think in term of units that have meaning to their work rather than a standardized measure of effect.

In many chapters of the text, the preferred approach is to begin with an estimate of raw measures. For the designs in this chapter that would be mean differences. For example, if designing a smoking cessation program and comparing smokers who participated with those placed on a waiting list for the program, determining a meaningful level of effectiveness for the program would be easier to accomplish when focusing on mean differences (e.g., Wilkinson & Task Force on Statistical Inference, 1999). In this case, we might determine that a difference of 10 cigarettes per day (half a pack) would be the minimal level of effectiveness required to term the approach successful. This approach does not preclude use of effect sizes; rather it encourages a focus on units relevant to the particular study.

This does not suggest that effect size estimates are not useful. Effect sizes are, of course, vitally important for understanding the context of differences. A difference of 10 cigarettes means less if your sample averages 50 cigarettes a day compared to a sample that averages 10 cigarettes a day. Similarly, a 10-cigarette difference is more meaningful if your samples produced a smaller standard deviation (e.g., 10 cigarettes) rather than a larger standard deviation (e.g., 40 cigarettes). Effect sizes are an important aspect of the context of power analysis, just not a great starting point for understanding meaningful differences between groups unless some standard is already established (e.g., effective smoking cessation programs produce a mean $d = 0.50$).

For those interested in beginning with effect sizes, see Example 3.5.

Example 3.1: Comparing Two Independent Groups

In collaboration with several colleagues, I helped to develop interactive computer-based tutorials for teaching core statistical concepts (see wise.cgu.edu for some of our work). An assessment of the effectiveness of one of the tutorials involved a quasi-experiment where a lab section of one course used a web-based tutorial and another lab section in the same course completed a standard assignment. Following each assignment, students completed a short exam on the topic.

To estimate the standard deviation, I examined previous grades on a 30-point exam used for several previous semesters in the same course. This allowed for a standard deviation estimate based on data from several hundred students. Scores from previous courses yielded a standard deviation of around 5.0. There was no reason to expect different standard deviations for the two groups, so I estimated σ_p at 5.0. The average score on the exam is 20.

Determining a meaningful effect involved comparing the work involved in implementing the new tutorial assignment to improved student outcomes. One practical consideration was the time required for instructors to implement the tutorial in courses. Implementation of the tutorial assignment involves several hours of work. Instructors would need to complete the tutorial assignment, work through common mistakes, anticipate student questions, and explore the interactive elements of the assignment to familiarize themselves with the capabilities of the instrument. In addition, because of the unfamiliar tutorial interface, the instructor would likely spend more time on student questions than for other types of assignments.

Given these new challenges, how large an effect would justify using the computer tutorial as a replacement for a more standard assignment? That is, how much improvement would convince instructors that the technique was worth their time? Based on previous experiences with the exam and what most students would accept as a "meaningful" improvement, I judged scores would have to improve performance by at least 2 points on the 30-point exam to justify the extra effort.

Calculation Examples (Approximate Power)

Given these criteria, we examine the sample size required to find a mean difference ($\mu_1 - \mu_2$) of 2 points with a standard deviation (σ_p) of 5. The following calculation shows this corresponds to $d = 0.40$.

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_p}$$

The next step is calculation of the approximate power for 80 students per group ($n = 80$ is an arbitrary value used to demonstrate calculations). Calculation of the

NCP yields $\delta = 2.53$. This approach requires the values for t above which the null is rejected (i.e., the critical value for t). For this example, with $n = 80$ per group, we have $df = 158$ with for $\alpha = .05$, two-tailed, corresponding to a $t_{critical}$ value of 1.98. The value of $\delta = 2.53$ is then compared to the $t_{critical}$ value of 1.98 using the formula that follows. The approximation approach for this independent samples example is appropriate to paired designs as well.

$$\delta = d\sqrt{\frac{n_j}{2}} = 0.40\sqrt{\frac{80}{2}} = 2.53$$

$$t_{power} = t_{critical} - \delta = 1.98 - 2.53 = -0.55$$

The calculation yields $t_{power} = -0.55$. This value reflects the point on the alternative distribution above which we reject the null hypothesis. This calculation does nothing more than convert a point on the null distribution to a point on the alternative distribution. Figure 3.1 represents these distribution points.[4] In this figure, there are two distributions. The one on the left is the null distribution. The x-axis represents scores on this distribution and range from –2.95 to +6.5. The $t_{critical}$ value of 1.98 is reflected on this axis. The value, δ represents the distance between the centers of the distributions, which is 2.53. As seen on the graph, 2.53 is the center

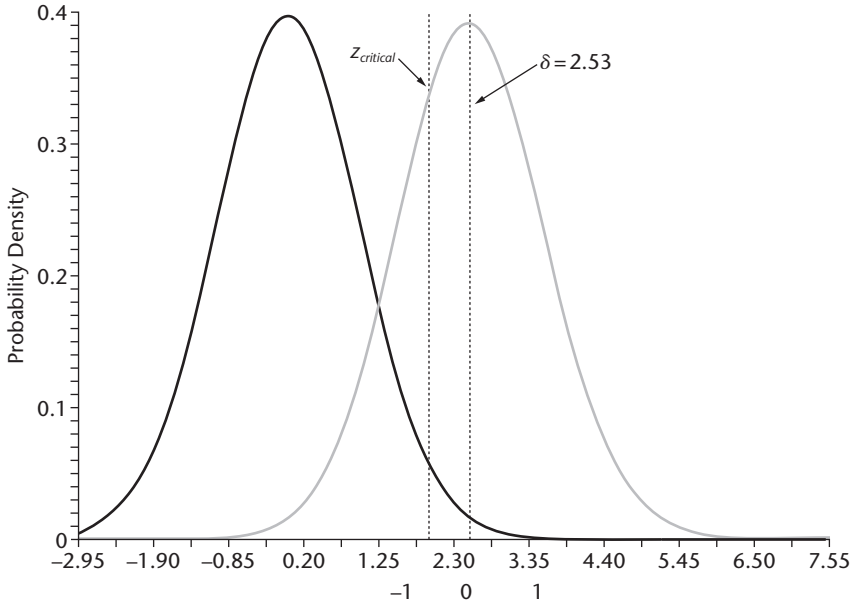

FIGURE 3.1 Demonstration of Noncentrality Parameter and Power.

Note
Distribution on the left is null. Distribution on the right is the alternative distribution.

of the distribution on the top *x*-axis. The area on the alterative distribution (the one on the right) that falls above the $t_{critical}$ value represents power. From the graphs here, we can see that this is a good chunk of the area, roughly about three-quarters.

The t_{power} value is not particularly meaningful for interpretation but it is necessary for calculations of power. One way to think of t_{power} is as a value that reflects a translation of the *x*-axis, which represents scores on the null distribution to an axis that represents scores on the alternative distribution. Conceptually, this would involve moving the axis over such that 0 now fell at the center of the alterative distribution. Figure 3.1 represents this with the values below the original *x*-axis (–1 to +1 in bold). Here the center of the alternative distribution would become zero. The t_{power} value of –0.55 is simply where the $t_{critcal}$ value would fall on the new *x*-axis.

Figure 3.1 also demonstrates an important difference between central and noncentral distributions. The null distribution is a central *t*-distribution. This distribution is a function of degrees of freedom and is symmetrical. The alternative distribution is a noncentral *t*-distribution. This distribution is a function of degrees of freedom and effect size. Note that the noncentral distribution is not completely symmetrical. Although it is subtle, if you look closely, you can see a slight positive skew to the distribution.

As in Chapter 2, we can approximate power using $t_{power} = -0.55$. The R command below takes *t* and *df* and yields the area above *t*. The command yields a value of .708 for power.

```
1-pt(–0.55, 158)
```

Calculation Examples (Exact Power)

R also provides an exact calculation of power. The calculation is based on noncentral distributions, accomplished using the R command below. The command requires the $t_{critical}$ value (1.98), *df* (158), and δ (2.53). The "1–" value remains regardless of the value of t_{power}.

```
1-pt(1.98, 158, 2.53)
```

Using this strategy, the calculation yields Power = .709. This appears consistent with Figure 3.1 that shows the power region as approximately three-quarters of the alternative distribution (the area to the right of $t_{critical} = 1.98$). Note that in this case, the approximation technique and the exact technique yield similar but not exactly equivalent results. This is because we have a relatively large sample size.

Solving for Sample Size

If we want to find Power = .80 (or some other value), we can work backward and rearrange some of the formulae. First, find the *t*-value above which 80% of

the distribution falls. The exact value the *t*-distribution changes depending on degrees of freedom. Practically however, this value changes very little as degrees of freedom rise above 10. For example $df = 10$, $t = -0.88$ is the point at or above which 80% of the distribution falls. For dfs $= 50$ and 1000 the corresponding *t*-values are –0.85 and –0.84, respectively. Of course, most designs involving two groups have $df > 10$ but as an approximation strategy, using $df = 10$ produces a reasonable estimate.

Next, using Formula 3.10, find δ. Then plug δ and $d = 0.40$ (from the calculation examples for approximate power) into Formula 3.11. The calculation indicates a sample size of 99 per group (n_j reflects the sample size per group, not the overall sample size). Also, recall that this technique uses an approximate rather than an exact technique.

$$\delta = t_{critical} - t_{power} \tag{3.10}$$

$$\delta = 1.98 - (-0.84) = 2.82$$

$$n_j = 2\left(\frac{\delta}{d}\right)^2 \tag{3.11}$$

$$n_j = 2\left(\frac{2.82}{0.40}\right)^2 = 99$$

Example 3.2: Power for Independent Samples *t* using R

The calculations for Example 3.1 are straightforward, but can be easily accomplished using R. Tables 3.1 provides R commands using the indt function. The general form of the function is as follows:

indt(m1, m2, s1, s2, n1, n2, alpha)

The values for *m*, *s*, and *n* refer to means, standard deviations, and sample size for each group. Alpha is set to default to .05.

Table 3.1 demonstrates use of this function for the values in Example 3.1. In the code, I entered means of 20 and 22 to reflect the two groups. This combined with the standard deviation of 5.0 produces the effect size of $d = 0.40$ used in the example. A second example in the table increased sample size until reaching the desired level of power (with $n = 99$ per group). For most analyses, you

TABLE 3.1 Power for Independent Samples *t*-test Using indt Function

```
indt(m1=22, m2=20, s1=5, s2=5, n1=80, n2=80)
## [1] "Equal Variance Power for n1 = 80, n2 = 80 = 0.71"
indt(m1=22, m2=20, s1=5, s2=5, n1=99, n2=99)
## [1] "Equal Variance Power for n1 = 99, n2 = 99 = 0.8"
```

will have to plug in a series of sample size values until you find the desired power level. See Chapter 11 for approaches (e.g., loops) for strategies yielding a series of estimates.

Example 3.3: Paired *t*-test

Instead of examining score improvement with a treatment group, a separate study examined improvement following student use of a technique where they took exams online and received immediate feedback on answers. Later in the semester, the students answered the same 30 items as part of a larger exam, allowing for a comparison of scores on the items. As the computer-based technique required programming of exams and feedback answers independently, application of the technique to new classes is time consuming, suggesting that a strong justification for the procedure would have to be present to make it worthwhile. Because the procedure is time consuming, it will only be considered effective if participants improve by 5 points on a subsequent exam on the same topics. As before, $\sigma = 5.0$.

This test requires an estimate of the correlation between the two exams. This was not necessary for the independent sample example, as participants were not tested twice. Estimating the correlation requires consideration of several factors. Correlations between measures in pre-post designs are often large (e.g., .70 or higher is not surprising). For research using reliable instruments or standardized measures there is often considerable information on scale reliability that informs this estimate. If there are no data addressing test–retest reliability, it is best to take a conservative approach to estimating the correlation between measures (ρ). I recommend using $\rho = .50$ in these situations. Larger correlations increase the effect size through reduction of the standard deviation of the differences whereas correlations below $\rho = .50$ reduce the effect size.

Calculation Examples

This example uses $n = 25$. As with most calculation examples, this is an arbitrary sample size used to demonstrate the calculation. In the example, notice that the standard deviation of the differences ($\sigma_{x_1 - x_2}$) is equal to the pooled standard deviation. That is because $\rho = .50$. Note that with $\rho = .70$, $\sigma_{x_1 - x_2} = 3.87$ and $d = 1.29$. With $\rho = .30$, $\sigma_{x_1 - x_2} = 5.92$ and $d = 0.84$.

$$\sigma_{x_1 - x_2} = \sigma_p \sqrt{2(1-\rho)} = 5.0\sqrt{2(1-0.50)} = 5\sqrt{1} = 5$$

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_p}$$

$$\delta = d\sqrt{n} = 1\sqrt{25} = 5$$

The NCP uses the overall sample size, since the same people are in each group. Starting with an estimate of 25 students participating in the study, we can calculate power.

Taking this value to a computer protocol for calculating noncentral probabilities, a sample of 25 students, $df = 24$ and *t*-critical value for a two-tailed test with $\alpha = .05$ of 2.06, yields Power = .998. Thus, a sample of 25 students provides considerable power to find differences of the size of interest if they exist. Alternatively, we could reduce the sample size and retain a decent level of power.

Example 3.4: Power for Paired *t* using R

The function pairt, demonstrated in Table 3.2, requires input that differs slightly from the independent samples test. Specifically, we enter a single standard deviation, an overall sample size, and the correlation between measures. The format of the function is as follows.

pairt(m1, m2, s, n, r, alpha)

The values for *m* refer to the two means, *s* to the pooled standard deviation, *n* for sample size, and *r* to the correlation. Alpha is set to default to .05. Table 3.2 demonstrates use of this function for the values in Example 3.3.

Given the power for 25 participation, it is reasonable to reduce the sample size if desired. The second example in Table 3.2 the code and output for a sample of 13 students. With roughly half of the initially proposed sample, we retain excellent power (.91). Therefore, in this case, substantially reducing sample size retains adequate power.

Example 3.5: Power from Effect Size Estimate

The tfromd function takes input of an effect size and outputs power for a range of sample sizes. The format of the function is:

tfromd(d, nlow, nhigh, alpha, test, tails, by)

The value *d* refers to the effect size. nlow and nhigh establish a range of sample sizes for estimating power. Alpha defaults to .05, enter a different value if desired (e.g., alpha = .01). The value test defaults to independent, use test = "P" for paired.

TABLE 3.2 Power for Paired Samples *t*-test using pairt Function

```
pairt(m1=25, m2=20, s=5, n=25, r=.5)
## [1] "Power for n = 25 is 0.998"
pairt(m1=25, m2=20, s=5, n=13, r=.5)
## [1] "Power for n = 13 is 0.911"
```

TABLE 3.3 Power using tfromd Function

```
tfromd(d=.2, nlow=10, nhigh=200, by=10, test="P")
## [1] "Power for total n of (Paired)  10 = 0.092"
## [1] "Power for total n of (Paired)  20 = 0.137"
## [1] "Power for total n of (Paired)  30 = 0.186"
## [1] "Power for total n of (Paired)  40 = 0.235"
## [1] "Power for total n of (Paired)  50 = 0.284"
## [1] "Power for total n of (Paired)  60 = 0.332"
## [1] "Power for total n of (Paired)  70 = 0.379"
## [1] "Power for total n of (Paired)  80 = 0.424"
## [1] "Power for total n of (Paired)  90 = 0.467"
## [1] "Power for total n of (Paired) 100 = 0.508"
## [1] "Power for total n of (Paired) 110 = 0.547"
## [1] "Power for total n of (Paired) 120 = 0.584"
## [1] "Power for total n of (Paired) 130 = 0.619"
## [1] "Power for total n of (Paired) 140 = 0.652"
## [1] "Power for total n of (Paired) 150 = 0.682"
## [1] "Power for total n of (Paired) 160 = 0.71"
## [1] "Power for total n of (Paired) 170 = 0.737"
## [1] "Power for total n of (Paired) 180 = 0.761"
## [1] "Power for total n of (Paired) 190 = 0.783"
## [1] "Power for total n of (Paired) 200 = 0.804"
```

Tails defaults to 2, use tails = 1 if desired. The value by sets the sample size estimates. The default is 1. For example, if nlow = 10 and nhigh = 14, with by = 1, the code produces estimates for power for 10, 11, 12, 13, and 14 participants. With by = 2, the code produces estimates for power for 10, 12, and 14.

Table 3.3 provides a paired *t*-test example using $d = 0.20$ and sample sizes ranging from 10 to 200.

Dealing with Unequal Variances, Unequal Sample Sizes, and Violation of Assumptions

The independent samples procedures presented earlier in the chapter assumed homogeneity of variances and employed calculations appropriate for homogenous variances and equal sample sizes. Heterogeneous variances and unequal sample sizes influence power for designs with independent samples so careful consideration of these issues may help the researcher avoid disappointment after study completion.

Homogeneity of Variance

The independent samples *t*-test assumes that the variances for the two groups are roughly equal. Most statistical packages provide output for analyses both

assuming and not assuming homogeneity. Some sources suggest that equal variance estimates are fine so long as the sample sizes are relatively equal (no more than a 4:1 ratio between the largest and smallest) and the largest variance is no more than 10 times the smallest variance. If the ratio of largest to smallest variance exceeds 10:1 or sample sizes exceeded a 4:1 then unequal variance estimates are preferred (Tabachnick & Fidell, 2007b). However, recent work suggests that alpha error inflates substantially with variance ratios of greater than 1.5:1 (Blanca, Alarcón, Arnau, Bono, & Bendayan, 2018). Others suggest using the unequal variances estimate as a default as such estimates provide better alpha error control and do not lose robustness when variances are equal (Delacre, Lakens, & Leys, 2017).

Many statistical packages include adjustments to degrees of freedom and standard error that account for violations of the heterogeneity of variance assumption. Formula 3.12 shows this adjustment, with a reduction in the degrees of freedom driven by the level of inequality between variances. I refer to this as $df_{unequal}$. This adjustment reduces the degrees of freedom to account for Type I error inflation resulting from unequal variances. Since the adjustment involves a reduction in degrees of freedom, the *t*-test probability for the unequal variances approach will be larger (usually) than the probability obtained using no adjustment. Of course, tests that make rejection criteria more stringent, result in a loss of power.

$$df_{unequal} = \frac{\left(\left(s_1^2 / n_1\right) + \left(s_2^2 / n_2\right)\right)^2}{\frac{\left(s_1^2 / n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2 / n_2\right)^2}{n_2 - 1}} \tag{3.12}$$

Transformation of data addresses many heterogeneity issues effectively. Heterogeneity often occurs because of non-normality. Transformations that return data to normality often address this problem adequately (see Tabachnick & Fidell, 2007a).

Heterogeneity and non-normality, in addition to influencing power through the degrees of freedom adjustment, often reduce the size of observed effects. To demonstrate this, Table 3.4 presents data for a treatment and a control group and Table 3.5 presents the summaries of raw and transformed analyses. The values in the first and second columns of Table 3.4 reflect raw data and the two rightmost columns reflect data subjected to a logarithmic transformation. In the raw data, the treatment group variance is over 100 times the size of the control group's variance (listed as variance ratio in the table) and the dependent variable (dv) shows large values for skew and kurtosis. A general strategy for evaluating the skew and kurtosis is to divide those values by their corresponding standard errors, with ratios of less than 3.0 indicating a roughly normal distribution (again see Tabachnick & Fidell, 2007b). Using these criteria, we see very large ratios

(listed as Skew Ratio and Kurtosis Ratio in Table 3.5). The raw data do not conform to the assumptions of the *t*-test. A test based on these data reduces power in two manners. First, the effect sizes differ considerably for the situation where we failed to meet assumptions ($d = 0.53$) compared to the transformed data analyses ($d = 1.52$). Second, if using the adjustment for unequal variances, $df_{unequal}$ would be smaller than the unadjusted *df* (21.3 vs. 42).

As demonstrated in Table 3.5, violating assumptions often affects power considerably. Techniques to address assumption violations are useful as a data analytic tool; however, it is also useful for power analyses to take homogeneity of variance issues into account when violations are expected. For example, if previous work with a scale regularly produced non-normal data then it is a good bet that future uses of the scale will do the same.

One strategy to address violations is to conduct a *t*-test that uses estimates appropriate for unequal variances. This strategy adjusts the degrees of freedom

TABLE 3.4 Demonstrating the Impact of Violation of Assumptions on Power

	Raw Scores		*Log Transformed*	
	Control	*Treatment*	*Control*	*Treatment*
	33	1811	1.53	3.26
	3200	441	3.51	2.65
	10	1081	1.04	3.03
	0	706	0.00	2.85
	0	730	0.00	2.86
	5	444	0.74	2.65
	328	715	2.52	2.85
	10,000	1968	4.00	3.29
	500	19,898	2.70	4.30
	26	21,331	1.43	4.33
	23	526	1.38	2.72
	656	669	2.82	2.83
	4	684	0.65	2.84
	10	12,503	1.03	4.10
	301	2685	2.48	3.43
	820	1632	2.91	3.21
	500	5986	2.70	3.78
	492	602	2.69	2.78
	3937	125,600	3.60	5.10
	13	3734	1.15	3.57
	19	20,121	1.30	4.30
	500	15,212	2.70	4.18
M	972	10,867	1.95	3.40
s	2260	26,629	1.16	0.70
s^2	5,107,600	709,103,641	1.35	0.49

TABLE 3.5 Summary Statistics for Raw and Transformed Data

	Raw Data	*Transformed*
s^2 Ratio	138.8	2.8
Skew	5.79	–0.48
SE Skew	0.36	0.36
Skew Ratio	16.1	1.3
Kurtosis	36.07	–0.16
SE Kurtosis	0.70	0.70
Kurtosis Ratio	51.5	0.2
d	0.53	1.52
Power	.40	1.0

($df_{unequal}$), as shown in Formula 3.12, for what is usually a more conservative test. The R code in Example 3.6 provides power analysis for adjusted and unadjusted tests. If you expect groups to exhibit even moderately unequal variances, use whichever power analysis suggests a larger sample size.

Unequal Sample Sizes

The formulae examined previously assumed equal sample sizes across groups. Practically, studies that lack random assignment make equal sample size requirements challenging. For most formulae, the harmonic mean of the two sample sizes as expressed in Formula 3.13 provides a good substitute. Simply place the harmonic mean in the formula for the noncentrality parameter (3.14).

$$n_{harmonic} = \frac{2n_1n_2}{n_1 + n_2} \tag{3.13}$$

$$\delta = d\sqrt{\frac{n_{harmonic}}{2}} \tag{3.14}$$

If sample sizes are not equal, equal and unequal variance approaches use different estimation methods for the standard error of the difference between mean, meaning that the denominator of the *t*-test changes depending on the approach. It is possible for unequal variance adjustments to produce tests that have more power than those with equal variances. However, this only occurs when you have unequal sample sizes and your largest group is the group with the larger variance.

Designing for Unequal Variances or Unequal Sample Sizes

If pretests or previous work suggest that control and treatment groups produce different variances, then it is best to address these issues in the design stage. If

information exists from previous work or pretesting then use those values to inform standard deviation estimates. If that information is not available but you do want to design for unequal variances and have only an estimate of the overall variance (or standard deviation), create two variances that pool to the overall variance with the constraint that one is much larger than the other. For example, to create two variances where one is 10 times the size of the other with a pooled variance expected to be 22, set the first group's variance as 40 and the second group as 4.

Designs that incorporate unequal sample sizes are a useful strategy for increasing power when one group is involved in an expensive treatment or reflects a hard-to-reach population. In these cases, assigning more participants to the cheaper of your treatments and sampling more from easier-to-obtain groups increases power (Lipsey, 1990; see Chapter 11 for more discussion of this approach).

Example 3.6: Unequal Variances and Unequal Sample Sizes

This example deals with legitimately unequal variances, that is, variances that are unequal because of differences between the groups rather than non-normal distributions. Differences of this nature do not respond to transformations and may be a product of the research design.

A few years back I developed a reaction time based on attitudes toward hate crimes. One phase of the study compared the reactions of gay and heterosexual men to the stimulus materials. As no similar measures existed, one approach to establishing validity of the instrument involved comparing the responses of the two groups, with the expectation that the gay men held attitudes that are more negative. At the time there was not a large, visible gay male population on our campus, so lab members recruited participants from the local community. The control group, heterosexual men, came from traditional sources (e.g., students in introductory psychology course).

To start the analysis, I estimated $\sigma = 1.0$ from previous work with the reaction time task with heterosexual men. The initial study designed for an effect size around $d = 0.60$ with the gay men indicating more negative responses to the stimuli than heterosexual men. The choice of $d = 0.60$ was determined by reference to other validity studies and was the smallest effect that allowed for a reasonable argument for the validity of the instrument. Using the procedures outlined in this chapter, these estimates produced a suggested sample size of $n = 45$ per group.

Following data collection, an initial look at the data revealed a mean difference that was larger than expected. There were, however, considerable differences in standard deviations between the groups. I expected both groups to demonstrate variances similar to an earlier group (heterosexual men). This assumption held for the heterosexual male sample who produced a standard

deviation of roughly 1.0. The gay male group showed a standard deviation of 4.0. This produced a variance that was 16 times larger than found for the heterosexual men. The larger than expected standard deviation for the comparison group created larger standard errors, a smaller than desirable *t*-statistic, and an effect size well under the desired value ($d \approx 0.30$). In short, the study failed to find differences between gay and heterosexual men, giving no support to claims of the instrument's validity.

In retrospect, some of these issues might have been avoided. The groups differed in obvious manners over and above sexual orientation. In particular, the gay men from the community were often older. During data collection, researcher assistants observed that these participants often took much longer to learn the computer task and sometimes could not respond to stimuli before response deadlines expired. The heterosexual (college-aged) men performed the task consistently, but the gay men (community sample) produced reaction time data that were all over the place.

A major source of error in the design was estimation of the standard deviation. It was a mistake to expect the comparison group to show the same standard deviation as the control group. Although this study represents a unique situation, it may be the case that in true experiments manipulations influence standard deviations as well as means. For example, a manipulation wherein one group solved problems while distracted and a control group solved problems without distraction, might produce more variability in the distracted conditions. Pretesting seems to be the only way to get a clear estimate of such differences, however, considering potential differences between groups that contribute to difference in variances is an important step in research design.

Calculations for Heterogeneity and Unequal Sample Size Adjustments

Using this example, I present calculations for a modification to the study that uses unequal sample sizes and addresses heterogeneity of variance. For the control group (heterosexual men), the standard deviation estimate remains $\sigma = 1.0$. For the experimental group (gay men), the new estimate is $\sigma = 4.0$. This design uses a sample of heterosexual men that is three times larger than the sample of gay men.

Initially, I was interested in differences of $d = 0.60$ or larger. However, a focus on effect size is a difficult starting point for analyses based on unequal variances. The effect size uses standard deviation units that undergo an unequal variance adjustment. Following the adjustment, your effect size may differ considerably from the one your starting effect size. The strategy I suggest focuses on an effect size without adjusting for unequal variances, and then using the mean differences (i.e., the numerator) associated with that value as the target difference. As seen later, this value may reflect a substantially smaller adjusted effect size.

The calculations that follow use an arbitrary starting sample size of 30 comparison group participants and 90 control group participants. The first step is to calculate the mean difference based on a pooled standard deviation estimate. Again, this simply provides some descriptive values for future calculations.

$$s_p = \sqrt{\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2}}$$

$$= \sqrt{\frac{26^2(19)+2.4^2(19)}{38}}$$

$$= \sqrt{\frac{12844+109.44}{38}}$$

$$= \sqrt{340.88}$$

$$= 18.46$$

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_p}$$

Next are calculations of standard deviation, degrees of freedom, and effect sizes that account for unequal variances. The primary issue here is calculation of a denominator for the *d* that does not use a pooled variance estimate. For lack of a better name, Formula 3.15 terms this $\sigma_{unequal}$. The effect size ($d_{unequal}$) is of interest as it is much smaller than the effect size based on a pooled standard deviation.

$$\sigma_{unequal} = \sqrt{\frac{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)}{\left(\frac{n_1+n_2}{n_1 n_2}\right)}}$$

$$\sigma_{unequal} = \sqrt{\frac{\left(\frac{4^2}{30}+\frac{1^2}{90}\right)}{\left(\frac{30+90}{30*90}\right)}} = 3.5 \qquad (3.15)$$

$$df_{unequal} = \frac{\left((4^2/30)+(1^2/90)\right)^2}{\frac{(4^2/30)^2}{30-1}+\frac{(1^2/90)^2}{90-1}} = 30.2$$

$$d_{unequal} = \frac{1.30}{3.5} = 0.37$$

TABLE 3.6 R Code and Output for Variance and Sample Adjusted Power

```
indt(m1=1.3, m2=0, s1=4, s2=1, n1=30, n2=90)
## [1] "Equal Variance Power for n1 = 30, n2 = 90 with
d = 0.601 = 0.807"
## [1] "Unequal Variance Power for n1 = 30, n2 = 90 with
d = 0.371 = 0.399"
indt(m1=1.3, m2=0, s1=4, s2=1, n1=78, n2=234)
## [1] "Equal Variance Power for n1 = 78, n2 = 234 with
d = 0.598 = 0.995"
## [1] "Unequal Variance Power for n1 = 78, n2 = 234 with
d = 0.371 = 0.801"
```

Next is the calculation of the harmonic mean of the sample sizes and the NCP. Recall that the harmonic mean is used when the sample sizes between groups differ.

$$n_{harmonic} = \frac{2n_1n_2}{n_1 + n_2} = \frac{2(30)(90)}{30 + 90} = 45$$

$$\delta = d\sqrt{\frac{n_{harmonic}}{2}} = 0.37\sqrt{\frac{45}{2}} = 1.76$$

Calculating power for $\delta = 1.76$ requires the *t*-critical value for two-tailed test with $\alpha = .05$ and $df = 30.2$. Recall that the $df = 30.2$ reflects the unequal variance adjustment to the degrees of freedom. Using R (the one line provided in the Exact Power section), power is .40, far less than the desired level of .80.

Independent Samples Commands for Unequal Variances and Unequal Sample Sizes

The tind function demonstrated in Example 3.2 carries out calculations for both equal and unequal variance estimates. In Table 3.6, note the differences in power depending on the estimation technique. For equal variances, a sample of 30 in the treatment and 90 in the control group yield power of .807 but the unequal variance power is only .40. Unequal variances tests do not achieve power of .80 until a sample of 78 in the treatment and 234 in the control group.

Additional Issues

There are several alternative approaches to the independent samples *t*-test for situations when data are non-normal or do not meet homogeneity assumptions. Developments in the use of bootstrapping techniques (a.k.a., re-sampling; e.g., Keselman, Othman, Wilcox, & Fradette, 2004) and other robust methods of analysis (e.g., Wilcox & Keselman, 2003) involve techniques that sometimes outperform the traditional *t*-test when assumptions fail. In general, these techniques do not perform as well when assumptions are met. Although bootstrapping and other

procedures are valuable, and often more powerful, alternatives to the procedures discussed in this chapter, there are not well-established conventions for the use of these procedures or power analysis strategies for these techniques.

Summary

This chapter examined power analysis for designs employing independent and paired *t*-tests. Independent samples designs require estimates of means and standard deviations (or just the effect size) to estimate power. Paired designs require means, standard deviations, and estimates of the correlation between measures. For independent samples, homogeneity of variance and unequal sample sizes influence power. Careful consideration of these issues establishes more accurate power estimates.

Notes

1. Although pilot studies provide reasonable estimates of variability, pilot testing requires substantially larger samples to establish accurate effective size estimates (Lakens & Evers, 2014).
2. Several variations on the standardized mean difference exist but the version presented here appears to be the most common form.
3. As with the d for the independent samples test, there are several variations of this statistic.
4. This graph was produced using ESCI software, see Chapter 11 for information on this outstanding visualization tool.

4

CORRELATIONS AND DIFFERENCES BETWEEN CORRELATIONS

This chapter examines power for tests of zero-order Pearson correlations and for tests of differences involving either independent or dependent correlations. Approaches to comparing dependent correlations are not widely presented in the behavioral sciences literature (i.e., these techniques do not appear in most statistics textbooks) so sections on those topics provide details on testing hypotheses as well as conducting power analysis.

Necessary Information

The tests covered in this chapter require specification of either a meaningful correlation (ρ) for the population or meaningful differences between population correlations. Procedures involving three or more correlations require specifications of correlations between all variables addressed in the procedure.

Factors Affecting Power

Correlations are measures of effect sizes, so larger correlations produce more power. For tests involving differences between correlations, the size of the difference to be detected, and if relevant, the correlation between the variables compared, influence power, with larger differences between predictors yielding more power and greater overlap between the predictors being compared yielding less power. In addition, differences between stronger correlations (e.g., .60 and .80) produce more power than tests of differences between smaller correlations (e.g., .20 and .40). For all tests, larger sample sizes and more liberal α increase power.

Zero-Order Correlation

Key Statistics

Addressing power for a zero-order correlation involves converting the correlation (ρ) to Cohen's d (Formula 4.1) and then computing a noncentrality parameter (NCP) (δ) from d (Formula 4.2). After computing these values, the analysis proceeds like the t-test procedures discussed in Chapter 3. Formula 4.2 for the calculation of δ uses n-2 in the numerator (degrees of freedom). There is no strong agreement on whether to use sample size (n) or degrees of freedom ($n-2$) for noncentrality and power calculations. My recommendation is to use degrees of freedom as this yields a more conservative test as this approach makes for a smaller NCP.

$$d = \frac{2\rho}{\sqrt{(1-\rho^2)}} \tag{4.1}$$

$$\delta = \frac{d\sqrt{n-2}}{2} \tag{4.2}$$

As discussed in previous chapters, computations of power for several tests in this chapter involve the noncentral t-distribution. Computer protocols allow us to calculate accurate values for power for those tests.

Example 4.1: Zero-order Correlations

An issue of particular interest in social psychology is the correlation between measures of attitudes and behaviors, behavioral intentions, or expectations. The example that follows reflects work in my laboratory examining how implicit and explicit attitudes differentially predict behavioral expectations of aggression. Expectations of aggression are important in that they relate to enactment of aggressive scripts that predict actual acts of aggression (Anderson & Bushman, 2002). One question asked in this example is whether a measure of implicitly held attitudes toward gay men predicts expectations of aggression.

An important starting point is to ask how large a correlation we want to be able to detect. The best question is not "how big the expected correlation is" but "how large the minimum meaningful correlation is." To establish context, one approach is to reference results from similar research. Meta-analytic results focusing on relationships between implicit attitudes and other behavior-relevant measures reported an average correlation of roughly .30 (Greenwald, Poehlman, Uhlmann, & Banaji, 2009). Of course, simply being the average correlation found across similar studies does not mean a correlation of .30 is practically meaningful. At this point, a reasonable question is whether this correlation is large enough to be of practical value. For research of this nature, this can be a slippery question. On the one hand, expectations of aggression do relate to aggressive behavior, and

even small reductions in aggressive behavior can be practically important. On the other hand, the research addresses measures potentially related to behavior rather than actual behaviors so there is a degree of separation between the dependent measure and actual acts of aggression. In addition, there is some evidence that the link between predictors and expectations often produces smaller effects than predictor–behavior relationships do (e.g., Greitemeyer, 2009). Therefore, it is reasonable to expect that the influence of attitudes on behaviors (particularly extreme behaviors like aggression) may be substantially smaller than the attitude–aggressive expectation link. For this reason, we settled on a correlation of .30, believing that this would provide some "cushion" for finding relationships between attitudes and actual behaviors in future research.

It is important to note that even small correlations can be meaningful. For example, as Rosenthal and Rubin (1982) discussed a correlation of .10 between a treatment and a life or death outcome corresponds to saving 10 more people's lives out of 100 than a treatment producing a correlation of .00 between treatment and outcome. Although a correlation of .10 corresponds to a "small" effect size (Cohen, 1988), clearly the treatment demonstrates a meaningful effect. Context is far more important than small, medium, and large labels.

Calculations

After determining that a meaningful correlation for the population in this example is $\rho = .30$, we can address power analysis. The example that follows begins with a sample size of 66 and a two-tailed test with $\alpha = .05$. The degrees of freedom for this test is n–2, yielding a critical value of $t(64) = 2.00$

$$d = \frac{2\rho}{\sqrt{(1-\rho^2)}} = \frac{2*0.30}{\sqrt{1-0.30^2}} = 0.629$$

$$\delta = \frac{d\sqrt{n-2}}{2} = \frac{0.629\sqrt{66-2}}{2} = 2.52$$

To find power for 66 participants, we can take the NCP to R. Using the line of code below, with a critical value of $t(64) = 2.00$ and $\delta = 2.52$, R computes Power = .70 (see the notes in Chapter 2 about using this approach).

R code: 1-pt(2.0, 64, 2.52)

R Function and Code

The code presented in Table 4.1 computes power for a range of sample sizes using the function corr. The form of the function is:

corr(r, nlow, nhigh, alpha, tails, by)

TABLE 4.1 R Code and Output for Zero-order Correlation Power Analysis

```
corr(r=.30, nlow=60, nhigh=100, by=2)
## [1] "Power for n of  60 = 0.6537"
## [1] "Power for n of  62 = 0.6689"
## [1] "Power for n of  64 = 0.6836"
## [1] "Power for n of  66 = 0.6978"
## [1] "Power for n of  68 = 0.7114"
## [1] "Power for n of  70 = 0.7246"
## [1] "Power for n of  72 = 0.7373"
## [1] "Power for n of  74 = 0.7495"
## [1] "Power for n of  76 = 0.7612"
## [1] "Power for n of  78 = 0.7724"
## [1] "Power for n of  80 = 0.7832"
## [1] "Power for n of  82 = 0.7936"
## [1] "Power for n of  84 = 0.8035"
## [1] "Power for n of  86 = 0.8131"
## [1] "Power for n of  88 = 0.8222"
## [1] "Power for n of  90 = 0.8309"
## [1] "Power for n of  92 = 0.8393"
## [1] "Power for n of  94 = 0.8473"
## [1] "Power for n of  96 = 0.8549"
## [1] "Power for n of  98 = 0.8622"
## [1] "Power for n of 100 = 0.8692"
```

The value *r* is the correlation. nlow, nhigh, and by set the range of values for calculation (as in Example 3.5). The values alpha and tails are the parameters of the significance test, the default values are .05 and 2, respectively.

As shown in the calculations section, use of 66 participants yield power of .70. Power of .80 requires a sample of 84 participants.

Comparing Two Independent Correlations

This test compares correlations drawn from independent populations. For example, this approach allows for tests involving meaningful differences among correlations between two variables measured for control group participants and the correlation between the same measures among experimental group participants in a between subjects design.

Formulae

The first step in this test is application of Fisher's transformation (Fisher, 1921) to the two expected population correlations. Formula 4.3 notes the converted value as z_ρ. Other sources represent the Fisher's transformed correlation as z_r or r'. Many statistics texts provide a table for this transformation, but it is not

difficult to compute with a hand calculator. The R package psych (Revelle, 2018) provides the command fishersz to compute the transformation as well.

$$z_\rho = 0.5 * \ln \frac{1+\rho}{1-\rho} \tag{4.3}$$

After calculating z_p for both correlations, take the difference between these values, noted as q and shown in Formula 4.4. This value is the effect size for the differences between the correlations.

$$q = \left| z_{\rho 1} - z_{\rho 2} \right| \tag{4.4}$$

Next is the calculation of z_δ. That value serves a role much like the NCP discussed in other sections. However, this is not a NCP because the normal distribution used for this test is a central rather than noncentral distribution. Calculation of z_δ requires a standard deviation as well. This value, shown in Formula 4.5, is a function of the sample sizes and serves as the denominator for Formula 4.6. The final calculation (Formula 4.7) finds the point on the alternative distribution that corresponds to the decision criteria. After computing z_{power}, take this value to a normal distribution table. The area above z_{power} reflects the proportion of sample results given the population correlations we specified that would allow for rejection of the null hypothesis. Since the normal distribution is a central distribution, it is possible to calculate power by hand accurately (see Chapters 1 and 2 for a discussion of this issue and for examples of the calculation techniques).

$$\sigma_q = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}} \tag{4.5}$$

$$z_\delta = \frac{q}{\sigma_q} \tag{4.6}$$

$$z_{power} = z_{critical} - z_\delta \tag{4.7}$$

Example 4.2: Comparing Independent Correlations

Extending Example 4.1, we predicted that participants assigned to different research conditions would show different levels of correlation between implicit attitudes and aggressive expectations. One condition involved priming of participants to think about their evaluations of gay men. In the second condition, there was no priming. We expected that in the first condition priming promoted deliberation, a situation usually linked to very small correlations between implicit attitudes and behaviors. The second condition promoted spontaneous processing, a situation linked to stronger implicit–behavior correlations. We predicted expectations to follow the same pattern seen for behaviors. For the first condition, we expected little relationship between attitudes and expectations, so we chose a small

correlation of .10 for this condition (noted ρ_1). The expected correlation for the second condition was .30 as before (noted as ρ_2).

Calculations

To determine power for detecting differences of this size, first convert both correlations using the Fisher transformation.

$$z_{\rho 1} = \frac{\ln(1+\rho) - \ln(1-\rho)}{2} = \frac{\ln(1+0.10) - \ln(1-0.10)}{2}$$
$$= \frac{0.095 - (-0.105)}{2} = 0.100$$

$$z_{\rho 2} = \frac{\ln(1+\rho) - \ln(1-\rho)}{2} = \frac{\ln(1+0.30) - \ln(1-0.30)}{2}$$
$$= \frac{0.262 - (-0.357)}{2} = 0.310$$

After calculating z_p for both correlations, calculate the effect size (q). Note that a difference between correlations of .20 as found produces different effect sizes depending on the correlation values. In this example $q = 0.21$. However, if the correlations were .60 and .80, then $q = 0.41$.

$$q = \left| z_{\rho 1} - z_{\rho 2} \right| = \left| 0.100 - 0.310 \right| = 0.210$$

Following the calculation of the effect size (q), calculate the standard deviation and z_δ. The example below computes power for a sample of 100 participants per condition.

$$\sigma_q = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}} = \sqrt{\frac{1}{100 - 3} + \frac{1}{100 - 3}} = 0.144$$

$$z_\delta = \frac{q}{\sigma_q} = \frac{0.210}{0.144} = 1.458$$

To obtain power for a one-tailed test with $\alpha = .05$ ($z = 1.645$) calculate z_{power} using the next formula. Since this test uses the normal distribution, it is not necessary to reference a noncentral distribution. To find power, use the normal distribution to determine the area corresponding to power. A sample of $n = 100$ per group yields power of .43. The value of .43 is simply the area above $z = 0.187$ on the normal distribution (see Chapter 2 for a graphical demonstration of this technique). For tests involving actual data (i.e., if you are using these formulae to perform a calculation rather than conduct a power analysis), the value calculated as z_δ is the test statistic.

$$z_{power} = z_{critical} - z_\delta = 1.645 - 1.458 = 0.187$$

TABLE 4.2 Code and Output for Comparing Two Independent Correlations

```
indcorr(r1=.3, r2=.1, nlow=200, nhigh=800, by=50, tails=1)
## [1] "Power for n of 200 = 0.4254"
## [1] "Power for n of 250 = 0.4956"
## [1] "Power for n of 300 = 0.559"
## [1] "Power for n of 350 = 0.616"
## [1] "Power for n of 400 = 0.6669"
## [1] "Power for n of 450 = 0.7119"
## [1] "Power for n of 500 = 0.7517"
## [1] "Power for n of 550 = 0.7866"
## [1] "Power for n of 600 = 0.8171"
## [1] "Power for n of 650 = 0.8436"
## [1] "Power for n of 700 = 0.8666"
## [1] "Power for n of 750 = 0.8865"
## [1] "Power for n of 800 = 0.9036"
```

R Code

The code in Table 4.2 performs calculations for a range of sample sizes (as shown in the output) using the indcorr function. The code is flexible, allowing for unequal sample sizes. For unequal sample sizes, specify the proportion of the sample in each group (as demonstrated in Chapter 3). The form of the function is:

indcorr(r1, r2, nlow, nhigh, propn1, alpha, tails, by)

The values r1 and r2 are the two correlations. The value propn1 sets the proportion of the sample in the first group with a default of .5 (for equal sample sizes). nlow, nhigh, and by again set the range of values for calculation. alpha and tails are the parameters of the significance test, with the same defaults as previously.

As shown in Table 4.2, using a one-tailed test with $\alpha = .05$, power reaches .80 at around $n = 300$ per group.

Comparing Two Dependent Correlations (One Variable in Common)

Dependent correlations are correlations that come from the same sample. For example, a researcher might measure the correlation between two predictors and one outcome variable and ask whether one of the predictors related more strongly to the outcome than the other. This section deals with situations where the correlations to be compared share a single variable in common (e.g., the same outcome variable but different predictors).

Key Statistics

Testing a difference between dependent correlations requires a more complex procedure than testing differences between independent correlations. Power depends on the correlation between the two predictors as well as their correlations with the shared outcome variable. For this test, it is important to establish accurate estimates of the correlations between predictors (ρ_{12}) as this value influences calculation of the NCP considerably. The first step in these calculations is to average the two correlations of interest using Formula 4.8. Next, Formula 4.9 defines a value noted as ρ_{det}. This is the determinant of the correlation matrix. You can also complete this calculation using a matrix algebra calculator (there are many freely available online). Finally using Formula 4.10, calculate the NCP (δ) and then address power based on the noncentral *t*-distribution as in previous sections.

The approach detailed next comes from the work of Williams (1959) and performs better in terms of Type I errors than approaches based on the normal distribution (Hittner, May, & Silver, 2003).

$$\bar{\rho} = \frac{(\rho_{y1} + \rho_{y2})}{2} \tag{4.8}$$

$$\rho_{\text{det}} = 1 - \rho_{y1}^2 - \rho_{y2}^2 - \rho_{12}^2 + 2\rho_{y1}\rho_{y2}\rho_{12} \tag{4.9}$$

$$\delta = \left|\rho_{y1} - \rho_{y2}\right| \sqrt{\frac{(n-1)(1+\rho_{12})}{2(n-1)/\rho_{\text{det}}(n-3) + \bar{\rho}^2(1-\rho_{12})^3}} \tag{4.10}$$

Example 4.3: Comparing Dependent Correlations, One Variable in Common

Extending the previous examples, another issue of interest is whether certain measures of attitudes predict expectations better than others do. In particular, do implicitly held attitudes predict aggression better than explicitly stated attitudes?

Tests of this nature require information about how strongly the two variables of interest correlate with the dependent measure and how strongly the predictors correlate with each other (i.e., the information in a correlation matrix with all three variables). In Example 4.1, we determined that a meaningful relationship between implicit attitudes and expectations was .30. Based on previous work, I expected the correlation between explicit attitudes and expectations to be no more than .04. Although this was a very weak relationship, the present test asks specifically if the implicit–expectation relationship is stronger than the explicit–expectation relationship. This is a different question than whether one relationship is statistically significant while the other is not. The final estimate needed is the correlation between the implicit and explicit measures. These values vary widely in the literature and are often context-specific. An earlier study in the research lab using both measures found a correlation of .20. This

was consistent with the range of correlations found in other studies examining relationships between implicit and explicit attitudes for other socially sensitive topics. These results suggest that .20 is a reasonable estimate of the correlation between predictors.

Calculations

The three correlations of interest were .30 for the implicit (1)-intention (y) relationship, .04 for explicit (2)-intention (y), and .20 for implicit (1)-explicit (2). Formula 4.8 calculates the average of the two predictor–dv correlations. Next, Formula 4.9 produces the determinant of the correlation matrix.

$$\bar{\rho} = \frac{(\rho_{y1} + \rho_{y2})}{2} = \frac{0.04 + 0.30}{2} = 0.17$$

$$\begin{aligned}\rho_{\text{det}} &= 1 - \rho_{y1}^2 - \rho_{y2}^2 - \rho_{12}^2 + 2\rho_{y1}\rho_{y2}\rho_{12} \\ &= 1 - 0.30^2 - 0.04^2 - 0.20^2 + 2(0.30)(0.04)(0.20) \\ &= 0.8732\end{aligned}$$

After calculating the average and determinant, Formula 4.10 yields the NCP. This example used $n = 100$. The degrees of freedom for this test is $n - 3$ and the critical value for a two-tailed test with $\alpha = .05$ is $t = 1.98$. Modifying the line of R code for calculating power given the NCP (see the section on zero-order correlation) to read 1-pt(1.98, 97, 2.11) yields Power = .55. (Note: For tests involving actual data the value calculated as δ is the test statistic.)

$$\begin{aligned}\delta &= \left|\rho_{y1} - \rho_{y2}\right| \sqrt{\frac{(n-1)(1+\rho_{12})}{\left(2(n-1)/(n-3)\right)\rho_{\text{det}} + \bar{\rho}^2(1-\rho_{12})^3}} \\ &= \left|0.30 - 0.04\right| \sqrt{\frac{(100-1)(1+0.2)}{\left(2(100-1)/(100-3)\right)0.8732 + 0.17^2(1-0.2)^3}} \\ &= 2.11\end{aligned}$$

R Code

The R code in Table 4.3 performs this calculation for a range of values using the function depcorr1. The function takes the following form:

depcorr1(r1y, r2y, r12, nlow, nhigh, alpha, tails, by)

The values r1y is the correlation between the first predictor and the dv, r2y is the correlation between the second predictor and the dv, and r12 is the correlation between the predictors. The remaining values are the same as in previous functions in the chapter.

TABLE 4.3 Code and Output for Comparing Two Dependent Correlations (One Variable in Common)

```
depcorr1(r1y=.3, r2y=.04, r12=.2, nlow=100, nhigh=300, by=10,
tails=2)
## [1] "Power for n of 100 = 0.5529"
## [1] "Power for n of 110 = 0.5949"
## [1] "Power for n of 120 = 0.634"
## [1] "Power for n of 130 = 0.6702"
## [1] "Power for n of 140 = 0.7035"
## [1] "Power for n of 150 = 0.7341"
## [1] "Power for n of 160 = 0.762"
## [1] "Power for n of 170 = 0.7875"
## [1] "Power for n of 180 = 0.8105"
## [1] "Power for n of 190 = 0.8314"
## [1] "Power for n of 200 = 0.8503"
## [1] "Power for n of 210 = 0.8673"
## [1] "Power for n of 220 = 0.8825"
## [1] "Power for n of 230 = 0.8962"
## [1] "Power for n of 240 = 0.9084"
## [1] "Power for n of 250 = 0.9192"
## [1] "Power for n of 260 = 0.9289"
## [1] "Power for n of 270 = 0.9375"
## [1] "Power for n of 280 = 0.9452"
## [1] "Power for n of 290 = 0.9519"
## [1] "Power for n of 300 = 0.9579"
```

Consistent with earlier calculations, a sample of $n = 100$ produces power of .55. The output shows that power of 80% requires a sample between 170 and 180. The code in Table 4.3 may be modified to get an exact sample size corresponding to the desired level of power (e.g., change nlow to 170, nhigh to 180, and by to 1).

Comparing Two Dependent Correlations (No Variables in Common)

This test compares two correlations based on two separate pairs of variables when all four variables are measured on the same sample. A common application of this technique is for repeated measures test of correlations. For example, this test is appropriate for determining whether the strength of a correlation between two variables differs across conditions that included the same participants. Questions of this nature might also examine whether the same variables correlate in the same manner across situations or over time.

Key Statistics

For these tests, the four variables yield six correlations representing all possible pairs. Power calculations use all six of these correlations. Formulae 4.11–4.14 note

correlations with the subscripts 1, 2, *x*, and *y*. Numbers correspond to the first measurement situation and the letters to the second. The test compares ρ_{12} and ρ_{xy}.

Formula 4.11 averages the correlations of interest. Formula 4.12 derives the covariance for the difference between the Fisher transformed correlations noted here as $cov_{\rho s}$. The *q* statistic found in Formula 4.13 examines the differences between the correlations of interest and requires the Fisher's transformation on each correlation using Formulae 4.3. The z_δ value calls for sample size, covariance between the correlations, and *q*. This test uses the normal distribution, so we plug z_δ into Formula 4.7. The procedure comes from Steiger (1980) with a modification proposed by Silver and Dunlap (1987). This is one of several procedures recommended by Silver, Hittner, and May (2004). As before, for tests involving actual data, z_δ is the test statistic.

$$\bar{\rho} = \frac{(\rho_{12} + \rho_{xy})}{2} \tag{4.11}$$

$$\text{cov}_{\rho s} = \frac{\begin{array}{c} 0.5 * \Big(\Big[\big(\rho_{1x} - \rho_{12}\rho_{2x}\big)\big(\rho_{2y} - \rho_{2x}\rho_{xy}\big) \Big] + \Big[\big(\rho_{1y} - \rho_{1x}\rho_{xy}\big)\big(\rho_{2x} - \rho_{12}\rho_{1x}\big) \Big] \\ + \Big[\big(\rho_{1x} - \rho_{1y}\rho_{xy}\big)\big(\rho_{2y} - \rho_{12}\rho_{1y}\big) \Big] + \Big[\big(\rho_{1y} - \rho_{12}\rho_{2y}\big)\big(\rho_{2x} - \rho_{2y}\rho_{xy}\big) \Big] \Big) \end{array}}{\big(1 - \bar{\rho}^2\big)^2} \tag{4.12}$$

$$q = \left| z_{\rho_{12}} - z_{\rho_{xy}} \right| \tag{4.13}$$

$$z_\delta = \frac{q\sqrt{n-3}}{\sqrt{2 - 2\,\text{cov}_{\rho s}}} \tag{4.14}$$

Example 4.4: Comparing Dependent Correlations, No Variables in Common

An example of this approach comes from the work of a former student and his advisor (Davis & Henry, 2008). They examined how strongly two variables correlated when measured in the research laboratory compared to measures of the same variables provided by the same participants online. Specifically they addressed the correspondence of feelings toward African Americans and symbolic racism. The researchers predicted stronger correlations for data collected online compared to data collected in the laboratory. Using the author's work as a template for designing a larger-scale investigation, the following example addresses the sample size necessary for power of .80.[1]

Calculations

This test requires all of the pairwise correlations, represented in a matrix of correlations in Table 4.4. Numbered labels designate the first set (correlations in the research laboratory) and letter designate the second set (Internet sample).

TABLE 4.4 Correlations between Variables for Comparing Two Dependent Correlations (No Shared Variables)

	1	*2*	*x*
1			
2	.40		
x	.30	.45	
y	.10	.35	.70

The calculations first find the average correlation between the two concepts of interest.

$$\bar{\rho} = \frac{(\rho_{12} + \rho_{xy})}{2} \frac{0.40 + 0.70}{2} = 0.55$$

This test requires application of Fisher's transformation to both correlations, and then computation of the difference between the two to calculate the effect size (q). This example assumes a two-tailed test and so uses the absolute value of the difference. This corresponds to a test of the difference in magnitudes, disregarding sign. For a test that considers the direction of difference, use the signed difference rather than the absolute difference.

$$z_{\rho_{12}} = \frac{\ln(1+\rho) - \ln(1-\rho)}{2} = \frac{\ln(1+0.40) - \ln(1-0.40)}{2} = 0.424$$

$$z_{\rho_{xy}} = \frac{\ln(1+\rho) - \ln(1-\rho)}{2} = \frac{\ln(1+0.70) - \ln(1-0.70)}{2} = 0.867$$

$$q = \left| z_{\rho_{12}} - z_{\rho_{xy}} \right| = \left| 0.424 - 0.867 \right| = 0.443$$

The extensive calculation that follows is the covariance between the correlations.

$$\text{cov}_{\rho s} = \frac{\begin{array}{l} 0.5 * \left(\left[\left(\rho_{1x} - \rho_{12}\rho_{2x} \right) \left(\rho_{2y} - \rho_{2x}\rho_{xy} \right) \right] + \left[\left(\rho_{1y} - \rho_{1x}\rho_{xy} \right) \left(\rho_{2x} - \rho_{12}\rho_{1x} \right) \right] \right. \\ \left. + \left[\left(\rho_{1x} - \rho_{1y}\rho_{xy} \right) \left(\rho_{2y} - \rho_{12}\rho_{1y} \right) \right] + \left[\left(\rho_{1y} - \rho_{12}\rho_{2y} \right) \left(\rho_{2x} - \rho_{2y}\rho_{xy} \right) \right] \right) \end{array}}{\left(1 - \bar{\rho}^2 \right)^2}$$

$$= \frac{\begin{array}{l} 0.5 * \left(\left[\left(0.3 - 0.4 * 0.45 \right) \left(0.35 - 0.45 * 0.7 \right) \right] + \left[\left(0.1 - 0.3 * 0.7 \right) \left(0.45 - 0.4 * 0.3 \right) \right] \right. \\ \left. + \left[\left(0.3 - 0.1 * 0.7 \right) \left(0.35 - 0.4 * 0.1 \right) \right] + \left[\left(0.1 - 0.4 * 0.35 \right) \left(0.45 - 0.35 * 0.70 \right) \right] \right) \end{array}}{\left(1 - 0.55^2 \right)^2}$$

$$= \frac{0.5 * \left[(0.12)(0.035) + (-0.11)(0.33) + (0.23)(0.31) + (-0.04)(0.205) \right]}{(0.6975)^2}$$

$$= \frac{0.5 * (0.0042 - 0.0363 + 0.0713 - 0.0082)}{0.4865} = \frac{0.0155}{0.4865} = 0.0319$$

The final step involves computation of z_δ and then use of that value in conjunction with the $z_{critical}$ value (.05, two-tailed in this example) to find power. In this case, power corresponds to the area above 0.65 on the standardized normal distribution. This area (power) is .26.

$$z_\delta = \frac{q\sqrt{n-3}}{\sqrt{2-2\,\mathrm{cov}_{\rho s}}} \frac{0.443\sqrt{20-3}}{\sqrt{2-(2*0.0319)}} = \frac{1.827}{1.391} = 1.31$$

$$z_{power} = z_{critical} - z_\delta = 1.96 - 1.31 = 0.65$$

R Code

Table 4.5 presents R code and output for comparisons between dependent correlations with no variables in common using the depcorr0 function. The function takes the following form:

depcorr0(r12, rxy, r1x, r1y, r2x, r2y, nlow, nhigh, alpha, tails, by)

The *r* values reflect the various correlations represented in Table 4.4. The remaining values are the same as in previous functions in the chapter.

The output in Table 4.5 shows that a sample of about 80 participants yields power of .80. Power reaches .90 with a sample of roughly 110.

TABLE 4.5 Code and Output for Comparing Correlations between Variables for Comparing Two Dependent Correlations (No Shared Variables)

```
depcorr0(r12=.4, rxy=.7, r1x=.3, r1y=.1, r2x=.45, r2y=.35,
nlow=20, nhigh=200, by=10, tails=2)
## [1] "Power for n of  20 = 0.2593"
## [1] "Power for n of  30 = 0.3808"
## [1] "Power for n of  40 = 0.4918"
## [1] "Power for n of  50 = 0.5893"
## [1] "Power for n of  60 = 0.6726"
## [1] "Power for n of  70 = 0.7421"
## [1] "Power for n of  80 = 0.7989"
## [1] "Power for n of  90 = 0.8447"
## [1] "Power for n of 100 = 0.881"
## [1] "Power for n of 110 = 0.9096"
## [1] "Power for n of 120 = 0.9317"
## [1] "Power for n of 130 = 0.9488"
## [1] "Power for n of 140 = 0.9618"
## [1] "Power for n of 150 = 0.9717"
## [1] "Power for n of 160 = 0.9791"
## [1] "Power for n of 170 = 0.9846"
## [1] "Power for n of 180 = 0.9887"
## [1] "Power for n of 190 = 0.9918"
## [1] "Power for n of 200 = 0.994"
```

Note on Effect Sizes for Comparing Correlations

The formulae presented in Formulae 4.4, 4.10, and 4.13 are applicable to tests that examine the magnitude of the differences between correlations and tests involving both magnitude and direction. For tests involving magnitude (e.g., is one variable a stronger predictor than another?), enter positive correlations for the values compared in the test (i.e., the values inside the absolute value notation), regardless of direction. The reasoning for this is two correlations may be similarly predictive but in opposite directions. For a test that examines both magnitude and direction, enter each correlation with its direction.

For example, consider a test that examines differences between dependent correlations with $\rho_{12} = .30$ and $\rho_{xy} = -.20$. For a test focused on magnitude only, the Fisher transformed correlations, applied to the absolute value of each correlation yields, 0.31 and 0.20, respectively. This produces $q = 0.11$.

$$q = \left| z_{\rho 12} - z_{\rho xy} \right| = \left| 0.31 - 0.20 \right| = 0.11$$

Now consider a test focused on magnitude and direction, again with $\rho_{12} = .30$ and $\rho_{xy} = -.20$. For this test, the Fisher transformed correlations, applied to the raw value of each correlation yields, 0.31 and –0.20, respectively. This produces $q = 0.51$.

$$q = \left| z_{\rho 12} - z_{\rho xy} \right| = \left| 0.31 - (-0.20) \right| = 0.51$$

Clearly, the choice of approach influences the effect size and subsequent power. For the formulae presented in 4.4, 4.10, and 4.13, regardless of the type of test used, all correlations outside of the absolute value notation should be entered with their appropriate direction (i.e., the raw correlation).

Additional Issues

The procedures used for comparing correlations, particularly those for comparing dependent correlations are but one of several approaches to these techniques. There remains considerable disagreement regarding the best procedures for comparing dependent correlations (see Silver et al., 2004 and Wilcox & Tian, 2008 for summaries and comparisons of other approaches). As a general consideration, most procedures for addressing these questions diverge from expected Type I error rates when data are non-normal. As noted in the *t*-test chapter, transforming data to normality can be an important step in maintaining appropriate error rates. Some of the questions addressed in this chapter, in particular the comparison of two independent correlations, might also be tested as a regression interaction. Chapter 9 includes more information on this approach.

Summary

This chapter examined power for zero-order correlations and for several tests comparing correlations. For zero-order correlations, the primary information required is the size of the correlation to detect (i.e., how large is a meaningful correlation). For tests comparing differences between correlations, power analyses require the size of a meaningful difference between the correlations and accurate estimates of correlations between all other variables (e.g., the correlation between the predictor variables compared).

Note

1. I take some liberties with the authors' data to provide a simple example.

5

BETWEEN SUBJECTS ANOVA (ONE AND TWO FACTORS)

Between subjects Analysis of Variance (ANOVA) designs focus on approaches where researchers either assign participants to or sample from independent groups. These tests often include planned or post hoc comparisons to detect differences between pairs of means or to interpret interactions. This chapter examines power for main effects, interactions, and contrasts/post hoc tests. Examples include one and two factor ANOVA, planned and post hoc contrasts, and simple effects tests. Additional issues include discussions of power for detecting all effects compared to power for an individual effect and artificial dichotomization.

Necessary Information

A good starting point is determining meaningful patterns of means (μs) and an estimate of standard deviation (σ) for each factor level. When approaching factorial ANOVA designs, it is necessary to determine cell means as well. Also important is a clear understanding of which effects are of interest (omnibus or contrast) as this influences sample size planning decisions.

Factors Affecting Power

In addition to sample size and Type I error rate, larger differences between means and smaller standard deviations yield more power. Also relevant to power are decisions regarding follow-up tests such as those involving planned comparisons between means and simple effects tests to examine interactions. Some approaches make no adjustment for inflation of α whereas others use some form of α adjustment (e.g., Bonferroni). Any downward adjustment to α reduces power.

Omnibus Versus Contrast Power

This chapter examines power for both omnibus tests and tests involving planned contrasts and simple effects. Unless there is a firm theoretical reason for the omnibus *F* being the primary focus, power analyses should focus on contrasts corresponding to research hypotheses. For example, if you want to conclude that two groups both outperform a third, then design for adequate power for those contrasts rather than for the omnibus test. Similarly, when hypotheses address specific interaction patterns, simple effects power is often more central to the research hypotheses than power for the interaction effect.

Key Statistics

There are two effect size statistics used for ANOVA power calculations. Partial eta squared ($\eta^2_{partial}$) and f^2. Studies of statistical power often present *f and f*^{2}. The more commonly reported effect size statistic is $\eta^2_{partial}$. This value is termed partial η^2 because for designs with multiple factors the variance explained by all factors except the effect of interest is partialed out of the calculation. This distinction is not important for one factor designs, as there are no other effects to partial out of the equation. Formula 5.1 shows the relationship between $\eta^2_{partial}$ and f^2.

For the noncentrality parameter (NCP), some approaches use degrees of freedom for the error whereas others use sample size. This produces small differences in power estimates, particularly when sample sizes or effect size are small. For this reason, some of the results produced in this text differ slightly from those produced by programs such as G*Power. Calculations in this text use df_{error} as it is more conservative (see Formula 5.2). Each of the calculations in Formulae 5.1–5.3 depends on sample size so it is tricky to start with partial η^2 and design from there. I prefer a strategy that establishes meaningful differences between pairs of means and often a focus on power for contrasts rather than power for omnibus tests.

$$f^2 = \frac{\eta^2_{partial}}{1-\eta^2_{partial}} \tag{5.1}$$

For the omnibus ANOVA, tests calculate λ using Formula 5.2.

$$\lambda = f^2 df_{error} \tag{5.2}$$

$$\eta^2_{partial} = \frac{SS_{effect}}{SS_{effect} + SS_{error}}. \tag{5.3}$$

Contrast tests require calculation of a noncentrality value (δ) that reflects the differences between the weighted (noted with *c*) means in a particular comparison. For the contrasts, examples present both λ and δ (Formulae 5.4 and 5.5). Both are correct ways to represent the NCP.

$$\delta = \frac{\left|\sum c_i \mu_i\right|}{\sqrt{MS_{w/in}\left(\sum_{j=1}^{p} \frac{c_j^2}{n_j}\right)}} \tag{5.4}$$

$$\lambda = \delta^2 \tag{5.5}$$

One Factor ANOVA Formulae

Formulae 5.6–5.9 are for equal sample sizes. With unequal sample sizes, replace n_j with the harmonic mean of the sample sizes (see Formula 3.13). The value j reflects the number of levels of the factor, or when used as a subscript, a notation to perform the operation for each group. The value μ_t reflects the grand mean (the mean of all scores irrespective of group).

$$MS_{w/in} = \sigma_p^2 \tag{5.6}$$

$$SS_{w/in} = MS_{w/in} * (n - j) \tag{5.7}$$

$$MS_{bg} = n_j \frac{\sum (\mu_j - \mu_t)^2}{j - 1} \tag{5.8}$$

$$SS_{bg} = n_j \sum (\mu_j - \mu_t)^2 \tag{5.9}$$

Factorial ANOVA Formulae

Calculations for factorial ANOVA follow a similar logic, with cell means and sample size replacing the j term (see Formulae 5.10–5.20). For main effects, A and B note levels of the factor with the same subscripts (A and B) used as j was. The $MS_{w/in}$ is the same as shown in Formula 5.6.

$$SS_{w/in} = MS_{w/in} * (n - cells). \tag{5.10}$$

$$MS_{bg} = n_{cell} \frac{\sum (\mu_{cell} - \mu_t)^2}{cells - 1} \tag{5.11}$$

$$SS_{bg} = n_{cell} \sum (\mu_{cell} - \mu_t)^2 \tag{5.12}$$

$$SS_A = n_A \sum (\mu_A - \mu_t)^2 \tag{5.13}$$

$$MS_A = n_A \frac{\sum (\mu_A - \mu_t)^2}{A - 1} \tag{5.14}$$

$$SS_B = n_B \sum (\mu_B - \mu_t)^2 \tag{5.15}$$

$$MS_B = n_B \frac{\sum(\mu_B - \mu_t)^2}{B-1} \tag{5.16}$$

$$SS_{AxB} = SS_{BG} - SS_A - SS_B \tag{5.17}$$

$$MS_{AxB} = \frac{SS_{AxB}}{(A-1)(B-1)} \tag{5.18}$$

Simple Effects Formulae

Test involving simple effects use the same logic as in the previous section but examine the influence of one factor isolated at the level of another.

$$SS_{AatB_i} = n_{AatB_i} \sum(\mu_{AatB_i} - \mu_{B_i})^2 \tag{5.19}$$

$$MS_{AatB_i} = \frac{SS_{AatB_i}}{A-1} \tag{5.20}$$

Example 5.1: One Factor ANOVA

This example presents a design to examine the effectiveness of dorm room interventions to improve intergroup attitudes. Based on previous work, college students score, on average 80 (μ) with a standard deviation of 10 (σ) on an established attitude scale. There are three separate interventions with the opportunity to assign students randomly to one of the interventions or a control group. The first treatment is an inexpensive program that involves students rooming with a student of another ethnic group and existing curricular enhancements where the students take a one-unit course (1 hour per week) in their first semester on campus. The second treatment involves the same roommate pairing but develops a new curriculum for the one-unit class. This program would also be relatively inexpensive. The third treatment involves a roommate pairing with a more extensive (and expensive) program utilizing a three-unit course (3 hours per week) with structured intergroup experiences that involve both roommates.

In determining an effect size for our design, the primary question should address the sort of effect that would be meaningful. This question is complex, but for the present example, a cost–benefit approach is relevant. The first two treatments involve low-cost options whereas the third program involves an expensive approach. To justify the high cost, it would be reasonable to expect a better performance from the third program (i.e., a larger effect size). A previous large-scale study examining predictors such as roommate contact in dorm rooms found moderate changes in ethnic attitudes predicted by roommate experiences such as living with students from other ethnic groups ($d \approx 0.30$; Van Laar, Levin, Sinclair, & Sidanius, 2005).

Based on these effects, combined with the cost of the interventions, we might decide to design to detect small effects for the inexpensive program ($d = 0.20$) but larger effects ($d = 0.60$) for the more expensive program. These effects would correspond to the following mean values: Control group with no intervention, $\mu = 80$; Treatment 1, $\mu = 82$; Treatment 2, $\mu = 82$; and Treatment 3, $\mu = 86$. Using these values with the standard deviation noted and $n = 60$ per group allows for several calculations.

$$MS_{w/in} = \sigma_p^2 = 10^2 = 100$$

$$MS_{bg} = n_{cell} \frac{\sum (\mu_{cell} - \mu_t)^2}{cells - 1}$$

$$= 60 \frac{(80 - 82.5)^2 + (82 - 82.5)^2 + (82 - 82.5)^2 + (86 - 82.5)^2}{4 - 1}$$

$$= 60 * \frac{19}{3} = 380$$

$$SS_{w/in} = MS_{w/in} * (n - cells) = 100 * (240 - 4) = 23600$$

$$SS_{bg} = n_j \sum \left(\mu_j - \mu_t\right)^2$$

$$= 60 * \left[\left(80 - 82.5\right)^2 + \left(82 - 82.5\right)^2 + \left(82 - 82.5\right)^2 + \left(86 - 82.5\right)^2 \right]$$

$$= 60 * 19 = 1140$$

$$\eta^2_{partial} = \frac{SS_{effect}}{SS_{effect} + SS_{error}} = \frac{1140}{1140 + 23600} = 0.0461$$

$$f^2 = \frac{\eta^2_{partial}}{1 - \eta^2_{partial}} = \frac{0.0461}{1 - 0.0461} = 0.0483$$

$$\lambda = f^2 df_{error} = 0.0483 * 236 = 11.4$$

For the η^2 and f^2 calculations, it is best to use several extra decimal places to ensure calculation accuracy. This likely will not make a great deal of difference calculation of power, but rounding often produces results that differ slightly from other sources.

Calculation of power for ANOVA and other designs using the *F* distribution is complex and best left to a computer. However, an approximation technique exists. Formula 5.21 presents the unit normal approximation approach to calculating power. As in other chapters, I present approximate formulae and then computer approaches for obtaining exact power estimates. Formula 5.21 differs slightly from the version that appears in Cohen (1988) to produce *z*-statistics consistent with the approaches discussed in Chapter 2.

$$z_{power} = \frac{\sqrt{(2df_{denom} - 1)\frac{df_{num}F_{critical}}{df_{denom}}} - \sqrt{2(df_{num} + \lambda) - \frac{df_{num} + 2\lambda}{df_{num} + \lambda}}}{\sqrt{\frac{df_{num}F_{critical}}{df_{denom}} + \frac{df_{num} + 2\lambda}{df_{num} + \lambda}}} \tag{5.21}$$

$$z_{power} = \frac{\sqrt{(2*236 - 1)\frac{3*2.643}{236}} - \sqrt{2(3 + 11.4) - \frac{3 + 2(11.4)}{3 + 11.4}}}{\sqrt{\frac{3*2.643}{236} + \frac{3 + 2(11.4)}{3 + 11.4}}} = -0.90$$

From z_{power} we find an area of .817 (this is the area above $z = -0.90$ on the normal curve). An exact calculation is accomplished using λ in conjunction with following line of R code.

```
1-pf(F_Table, dfbg, dfwin, Lambda)
```

Using the values yields:

```
1-pf(2.643, 3, 236, 11.4)
```

Using this approach, Power = .812 (the .005 difference resulting from the approximation used Formula 5.21). This result suggests adequate power (>.80) for the omnibus test. However, detecting differences on the omnibus test may not be the primary effect of interest. For example, if our interest was whether the new programs outperform the older practices (no program and the current program), if the current program is better than no program, and whether one of the new programs is substantially better than the other, then planned comparisons rather than an omnibus test would be more relevant. The next section includes calculations for power for contrasts then provides R code for both the omnibus and contrast tests.

Example 5.2: One Factor ANOVA with Orthogonal Contrasts

The present study allows for a set of three orthogonal contrasts. One contrast that makes sense involves comparing the Control and Current procedures (Groups 1 and 2) to the New and Extended procedures (Groups 3 and 4). This contrast establishes whether the new programs differ from what the campus is currently doing. Contrast 2 is Control vs. Current and Contrast 3 compares the two new programs (New vs. New Extended). Calculating contrast values for each involves placing a weight on each mean as shown in Table 5.1 (for more on the contrast procedures, see Keppel, 1991 or Kirk, 1995). The weights in this table serve as the values of "*c*" in Formula 5.4.

TABLE 5.1 Contrast Weights (c) for One Factor ANOVA Example

Contrast	*Control*	*Current*	*New*	*New Extended*
1	1	1	–1	–1
2	1	–1	0	0
3	0	0	1	–1

I labeled each δ and λ with a subscript reflecting the contrast.

$$\delta_1 = \frac{\left|\sum c_i \mu_i\right|}{\sqrt{MS_{w/in}\left(\sum_{j=1}^{p} \frac{c_j^2}{n_j}\right)}}$$

$$= \frac{\left|(1*80)+(1*82)+(-1*82)+(-1*86)\right|}{\sqrt{100\left(\frac{1^2}{60}+\frac{1^2}{60}+\frac{-1^2}{60}+\frac{-1^2}{60}\right)}} = \frac{6}{2.582} = 2.324$$

$$\lambda_1 = 2.324^2 = 5.40$$

$$\delta_2 = \frac{\left|(1*80)+(-1*82)+(0*82)+(0*86)\right|}{\sqrt{100\left(\frac{1^2}{60}+\frac{1^2}{60}+\frac{0^2}{60}+\frac{0^2}{60}\right)}} = \frac{2}{1.826} = 1.095$$

$$\lambda_2 = 1.095^2 = 1.20$$

$$\delta_3 = \frac{\left|(0*80)+(0*82)+(1*82)+(-1*86)\right|}{\sqrt{100\left(\frac{1^2}{60}+\frac{1^2}{60}+\frac{0^2}{60}+\frac{0^2}{60}\right)}} = \frac{4}{1.826} = 2.191$$

$$\lambda_3 = 2.191^2 = 4.80$$

Next, take the values for δ to R, using the code that follows.

1-pt($t_{critical}$, df, δ)

For this test, degrees of freedom correspond to *df* error from the ANOVA (236 in this case). The critical value for two-tailed t at .05 is 1.97. Alternatively, the code presented earlier for F and Lambda produces the same result. For the first contrast, the code is:

1-pt(1.97, 236, 2.32)

This yields power of .64. This suggests that if the first contrast involves a research question of interest designing for adequate power (e.g., .80) requires a larger sample size.

TABLE 5.2 Power for $n = 60$ per group for Omnibus Test and Contrasts

	λ (δ)	t_{power}	*Power*
Omnibus *F*	11.40	–0.90	.81
Contrast 1	5.40 (2.32)	–0.36	.64
Contrast 2	1.20 (1.10)	0.86	.19
Contrast 3	4.80 (2.19)	–0.23	.59

Table 5.2 summarizes power for the remaining contrasts. Note that each contrast failed to produce power that approached the level of the omnibus test. This is not always the case, but outcomes like this are common enough that, unless there is a firm theoretical reason for omnibus *F* as the primary focus of the research, it is better to focus design efforts and power analyses on contrasts.

R Code for the One Factor ANOVA

Table 5.3 presents R code and output for completing the analyses detailed earlier. Note that the code requires the descriptive statistics but no calculations. The format of the functions are as follows:

anova1f_4(m1, m2, m3, m4, s1, s2, s3, s4, n1, n2, n3, n4, alpha)

anova1f_4c(m1, m2, m3, m4, s1, s2, s3, s4, n1, n2, n3, n4, alpha, c1, c2, c3, c4)

The values m1-m4, s1-s4, and n1-n4 reflect means, standard deviations, and sample size for each factor level, respectively. Alpha defaults to .05 if no value is entered. For the second function, the values c1-c4 reflect the contrast weights.

At this point, it is important to consider which hypotheses are of the most interest and design for optimal power on those specific contrasts. For example,

TABLE 5.3 R Code and Output Omnibus *F* and Contrasts ($n = 60$ per cell)

```
anova1f_4(m1=80, m2=82, m3=82, m4=86, s1=10, s2=10, s3=10,
s4=10, n1=60, n2=60, n3=60, n4=60)
## [1] "Power = 0.812"
anova1f_4c(m1=80, m2=82, m3=82, m4=86, s1=10, s2=10, s3=10,
s4=10, n1=60, n2=60, n3=60, n4=60, c1=1, c2=1, c3=-1, c4=-1,
alpha=.05)
## [1] "Power for Contrast = 0.638"
anova1f_4c(m1=80, m2=82, m3=82, m4=86, s1=10, s2=10, s3=10,
s4=10, n1=60, n2=60, n3=60, n4=60, c1=1, c2=-1, c3=-0, c4=0,
alpha=.05)
## [1] "Power for Contrast = 0.194"
anova1f_4c(m1=80, m2=82, m3=82, m4=86, s1=10, s2=10, s3=10,
s4=10, n1=60, n2=60, n3=60, n4=60, c1=0, c2=0, c3=1, c4=-1,
alpha=.05)
"Power for Contrast = 0.588"
```

TABLE 5.4 R Code and Output Contrasts ($n = 100$ per cell)

```
anova1f_4c(m1=80, m2=82, m3=82, m4=86, s1=10, s2=10, s3=10,
s4=10, n1=100, n2=100, n3=100, n4=100, c1=1, c2=1, c3=-1,
c4=-1, alpha=.05)
## [1] "Power for Contrast = 0.849"
anova1f_4c(m1=80, m2=82, m3=82, m4=86, s1=10, s2=10, s3=10,
s4=10, n1=100, n2=100, n3=100, n4=100, c1=1, c2=-1, c3=0,
c4=0, alpha=.05)
## [1] "Power for Contrast = 0.292"
anova1f_4c(m1=80, m2=82, m3=82, m4=86, s1=10, s2=10, s3=10,
s4=10, n1=100, n2=100, n3=100, n4=100, c1=0, c2=0, c3=1,
c4=-1, alpha=.05)
## [1] "Power for Contrast = 0.806"
```

a reasonable decision is to focus on power for Contrasts 1 and 3. These tests address whether the new procedures outperform the old and whether the extended new procedure outperforms the new program without extension. As before, we can modify the sample sizes in the R code in Table 5.3 and re-run the analysis until we find the optimal level of power. The code and output in Table 5.4 shows that 100 participants per cell produces power > .80 for Contrasts 1 and 3.

Polynomial Contrasts for One Factor ANOVA

Polynomial contrasts produce trend analyses. There are four levels to our factor, allowing for tests of the linear, quadratic, and cubic trends. Chapter 7 includes a more detailed discussion of trend analyses.

The pwr_anova1f_4c function allows for polynomial contrasts. For four factors, the linear contrasts are –3, –1, 1, 3. The quadratic contrast is 1, –1, –1, 1. The cubic contrast is –1, 3, –3, 1. See Kirk (1995), Keppel (1991), or perform an internet search on the term "polynomial contrasts" for codes for designs with different factor levels (see Table 5.5 for an example).

TABLE 5.5 R Code and Output for Polynomial Contrasts

```
anova1f_4c(m1=80, m2=82, m3=82, m4=86, s1=10, s2=10, s3=10,
s4=10, n1=60, n2=60, n3=60, n4=60, c1=-3, c2=-1, c3=1, c4=3,
alpha=.05)
## [1] "Power for Contrast = 0.874"
anova1f_4c(m1=80, m2=82, m3=82, m4=86, s1=10, s2=10, s3=10,
s4=10, n1=60, n2=60, n3=60, n4=60, c1=1, c2=-1, c3=-1, c4=1,
alpha=.05)
## [1] "Power for Contrast = 0.12"
anova1f_4c(m1=80, m2=82, m3=82, m4=86, s1=10, s2=10, s3=10,
s4=10, n1=60, n2=60, n3=60, n4=60, c1=-1, c2=3, c3=-3, c4=1,
alpha=.05)
## [1] "Power for Contrast = 0.179"
```

Comparisons among All Means

Although power analysis is an a priori venture, some research situations call for designs using conservative post hoc analyses. Conservative post hoc options (e.g., Tukey tests) involve comparisons between all pairs of means.

The R code (again using the anova1f_4c function) in Table 5.6 takes each mean, and assigns it as the comparison group for a simple contrast. This requires six tests to cover all the comparisons produced by the four factor levels.

The tests presented in Table 5.6 do not conduct tests such as the Tukey HSD. However, Bonferroni or Šidák adjustments provide a reasonable approximation. Formulae 5.22 and 5.23 detail these adjustments. With four groups, there are six comparisons between means. To adjust for tests using $\alpha = .05$, we end up with Bonferroni: $\alpha = .0083$ and Šidák: $\alpha = .0085$. Enter values from whichever test you plan to use in the code (the example uses Šidák).

$$\alpha_{Bonferroni} = \frac{\alpha}{c} \tag{5.22}$$

$$\alpha_{Sidak} = 1 - (1 - \alpha)^{1/c} \tag{5.23}$$

ANOVA with Two Factors

Power for effects involving multiple factors in ANOVA have been described elsewhere with a focus that begins with estimating the effect size associated with main effects and interactions. This approach is difficult as effect sizes for

TABLE 5.6 R Code and Output for All Pairwise Comparisons

```
anova1f_4c(m1=80, m2=82, m3=82, m4=86, s1=10, s2=10, s3=10, s4=10,
n1=60, n2=60, n3=60, n4=60, c1=1, c2=-1, c3=0, c4=0, alpha=.0085)
## [1] "Power for Contrast = 0.061"
anova1f_4c(m1=80, m2=82, m3=82, m4=86, s1=10, s2=10, s3=10, s4=10,
n1=60, n2=60, n3=60, n4=60, c1=1, c2=0, c3=-1, c4=0, alpha=.0085)
## [1] "Power for Contrast = 0.061"
anova1f_4c(m1=80, m2=82, m3=82, m4=86, s1=10, s2=10, s3=10, s4=10,
n1=60, n2=60, n3=60, n4=60, c1=1, c2=0, c3=0, c4=-1, alpha=.0085)
## [1] "Power for Contrast = 0.736"
anova1f_4c(m1=80, m2=82, m3=82, m4=86, s1=10, s2=10, s3=10, s4=10,
n1=60, n2=60, n3=60, n4=60, c1=0, c2=1, c3=-1, c4=0, alpha=.0085)
## [1] "Power for Contrast = 0.008"
anova1f_4c(m1=80, m2=82, m3=82, m4=86, s1=10, s2=10, s3=10, s4=10,
n1=60, n2=60, n3=60, n4=60, c1=0, c2=1, c3=0, c4=-1, alpha=.0085)
## [1] "Power for Contrast = 0.324"
anova1f_4c(m1=80, m2=82, m3=82, m4=86, s1=10, s2=10, s3=10, s4=10,
n1=60, n2=60, n3=60, n4=60, c1=0, c2=0, c3=1, c4=-1, alpha=.0085)
## [1] "Power for Contrast = 0.324"
```

interactions are neither intuitive nor specific to the pattern of effects of interest. For example, designing for an interaction with a partial η^2 of .03 provides no information about whether this is a meaningful result or the pattern of effects underlying the interaction.

Because of these concerns, I prefer an approach that generates a set of meaningful pattern of cell means. That is, what is the specific pattern of result that is of interest and what means correspond to the result? Determining meaningful means for cells requires considerable thought. Far more than simply designing for a certain effect size. However, this approach likely provides better power and sample size estimates.

Example 5.3: Two Factor ANOVA with Interactions

One of my long-standing areas of research involves examining attitudes toward affirmative action (AA). One project examined how policy features and the presence of policy justifications influence support for different forms of AA. There is considerable research on topics relevant to both justification and policy type. However, little is known about justifications for specific applications of AA as most work on justification examined attitudes toward AA in general and did not manipulate policy type and justification in the same study.

The study in this example uses a 2 (Policy Type: Recruitment of Applicants vs. Tiebreaker) by 2 (Justification: No Justification or Increased Diversity) design. The dependent measure was a four-item policy support scale used in a previous study, producing a standard deviation of 1.70. Table 5.7 details the expected cell and marginal means (an explanation of how I generated these values appears later). The primary hypothesis for the present study was that justifications influence evaluations of stronger policies like the tiebreaker policy wherein minority applicants received preference when their qualifications were equal to those of a non-minority applicant. However, justifications were not expected to influence policies wherein organizations made special outreach efforts to recruit minority applicants.

Several sources of information went into the determination of cell means. Results from a meta-analysis focusing on diversity justifications reported a correlation of .17 for differences between justified and not justified policies among studies examining attitudes toward AA in general (Harrison, Kravitz, Mayer, Leslie, & Lev-Arey, 2006). Analyses apply this effect size to the expected value for tiebreak (but not recruitment) conditions as previous work found individuals view tiebreakers as more typical of AA than is recruitment (Aberson, 2007). Converting r to d using Formula 4.1 yields $d = 0.35$. Given $\sigma = 1.7$, this corresponds to a difference of 0.60 between justified and not justified tiebreaker policies. Table 5.7 shows this difference in the row labeled "Tiebreak."

For policies presented without justifications, previous work in my lab found a difference in support for recruitment and tiebreaker policies that were presented

TABLE 5.7 Means for Factorial ANOVA Example

		Factor B		*M*
		No Justification (B1)	*Justified (B2)*	
Factor A	Recruit (A1)	0.85 (m1.1)	0.85 (m1.2)	0.85
	Tiebreak (A2)	0.0 (m2.1)	0.60 (m2.2)	0.30
	M	0.425	0.725	0.575

without justification of roughly $d = 0.50$ favoring recruitment approaches. Given $\sigma = 1.7$ (the standard deviation from a previous use of the scale), this corresponds to a difference of 0.85 between recruitment and tiebreaker policies. The column of Table 5.7 labeled "No justification" shows this difference.

The final cell mean is for the recruit-justified condition. Although there are no previous data to base this on, part of the interaction hypothesis was that justification would not influence evaluations of the recruitment policy. Thus, analyses set the means to show no difference between the recruitment policies between not justified and justified conditions.

Calculations

Power for the factorial ANOVA may be calculated (mostly) by hand using the approaches that follow. Although the primary hypothesis involves the interaction, the calculation approach addresses all of the effects (but presents power only for the interaction). Calculations began with an estimate of $n = 100$ per cell. This may seem like a very large sample but the study itself involved only a single page of measures that took participants roughly 2 minutes to complete so large samples sizes were not unreasonable.

$$MS_{w/in} = \sigma_p^2 = 1.7^2 = 2.89$$

$$SS_{w/in} = MS_{w/in} * (n - cells) = 2.89 * (400 - 4) = 1144.4$$

$$MS_{bg} = n_{cell} \frac{\sum (\mu_{cell} - \mu_t)^2}{cells - 1}$$

$$= 100 \frac{(0.85 - 0.575)^2 + (0.00 - 0.575)^2 + (0.85 - 0.575)^2 + (0.60 - 0.575)^2}{4 - 1}$$

$$= 100 \left(\frac{0.4825}{3} \right) = 16.08$$

$$SS_{bg} = n_{cell} \sum (\mu_{cell} - \mu_t)^2$$

$$= 100 \left[(0.85 - 0.575)^2 + (0.00 - 0.575)^2 + (0.85 - 0.575)^2 + (0.60 - 0.575)^2 \right]$$

$$= 100 (0.4825) = 48.25$$

$$MS_A = n_A \frac{\sum(\mu_A - \mu_t)^2}{A-1}$$
$$= 200 \times \frac{(0.425-0.575)^2 + (0.725-0.575)^2}{2-1} = 9.0$$

$$SS_A = n_A \sum\left(\mu_A - \mu_t\right)^2$$
$$= 200\left[\left(0.425-0.575\right)^2 + \left(0.725-0.575\right)^2\right] = 9.0$$

$$MS_B = n_B \frac{\sum\left(\mu_B - \mu_t\right)^2}{B-1}$$
$$= 200 \times \frac{\left(0.850-0.575\right)^2 + \left(0.300-0.575\right)^2}{2-1} = 30.25$$

$$SS_B = n_B \sum\left(\mu_B - \mu_t\right)^2$$
$$= 200\left[\left(0.850-0.575\right)^2 + \left(0.300-0.575\right)^2\right] = 30.25$$

$$SS_{AxB} = SS_{BG} - SS_A - SS_B = 48.25 - 9.0 - 30.25 = 9.0$$

$$MS_{AxB} = \frac{SS_{AxB}}{(A-1)(B-1)} = \frac{9.0}{(2-1)(2-1)} = 9.0$$

$$\eta^2_{partial} = \frac{SS_{effect}}{SS_{effect} + SS_{error}} = \frac{9}{9+1144.4} = 0.0078$$

$$f^2 = \frac{\eta^2_{partial}}{1-\eta^2_{partial}} = \frac{0.0078}{1-0.0078} = 0.0079$$

$$\lambda = f^2 df_{error} = 0.0079 \times 396 = 3.11$$

Taking $\lambda = 3.11$ to R (using the single line of code demonstrated earlier in the chapter) finds power of .42. A sample of $n = 100$ per cell is not adequate if we want power of .80.

R Code for Factorial ANOVA

Power calculations for this design use the anova2x2 function. The structure of the function is as follows:

> anova2x2(m1.1, m1.2, m2.1, m2.2, s1.1, s1.2, s2.1, s2.2, n1.1, n1.2, n2.1, n2.2, alpha, all)

The values for *m*, *s*, and *n* correspond to cell means (see the numbering in Table 5.7). alpha defaults to .05. I discuss the remaining item ("all") in the section on multiple effects.

TABLE 5.8 R Code and Output for Two Factor ANOVA

```
anova2x2(m1.1=0.85, m1.2=0.85, m2.1=0.00, m2.2=0.60,
s1.1=1.7, s1.2=1.7, s2.1=1.7, s2.2=1.7,
n1.1=100, n1.2=100, n2.1=100, n2.2=100,
alpha=.05)
## [1] "Power for Main Effect Factor A = 0.898"
## [1] "Power for Main Effect Factor B = 0.421"
## [1] "Power for Interaction AxB = 0.421"
anova2x2(m1.1=0.85, m1.2=0.85, m2.1=0.00, m2.2=0.60,
s1.1=1.7, s1.2=1.7, s2.1=1.7, s2.2=1.7,
n1.1=250, n1.2=250, n2.1=250, n2.2=250,
alpha=.05)
## [1] "Power for Main Effect Factor A = 0.999"
## [1] "Power for Main Effect Factor B = 0.796"
## [1] "Power for Interaction AxB = 0.796"
```

The R code in Table 5.8 completes all of the intermediate calculations and provides an estimate of power for the two main effects and the interaction using $n = 100$ per group. As in the earlier example, Power = .42 for the interaction. Also included in the table is an analysis that increases the sample size to $n = 250$. This analysis finds power of roughly .80 for the interaction.

Simple Effect Tests

Of primary interest when examining interactions are simple effects tests. Much as contrasts offer a more precise explanation of effects than do omnibus F tests, simple effects tests address specific aspects of the interaction by focusing on the influence of one factor at the levels of the other. For the current example, the hypothesis stated that justification makes a difference for tiebreaker policies but not for recruitment policies. To test these predictions, we can examine differences in support for tiebreaker policies between the justified or not justified conditions and then differences in support for recruitment policies between the justification conditions. Addressing power for simple effects tests is important as these tests often relate directly to hypotheses.

$$SS_{AatB_i} = n_{AatB_i} \sum \left(\mu_{AatB_i} - \mu_{B_i} \right)^2$$

$$SS_{Justify\_at\_Tiebreak} = 251\left(0.0 - 0.3\right)^2 + 251\left(0.6 - 0.3\right)^2 = 45.18$$

The following calculations yield $\lambda = 15.2$, for this value, power is .97, suggesting that the current design provides excellent power for detecting the simple effect of interest.

TABLE 5.9 R Code and Output for Simple Effects

```
anova2x2_se(m1.1=0.85, m1.2=0.85, m2.1=0.00, m2.2=0.60,
s1.1=1.7, s1.2=1.7, s2.1=1.7, s2.2=1.7, n1.1=250, n1.2=250,
n2.1=250, n2.2=250, alpha=.05)
## [1] "Simple Effect Comparing M = 0.85 and 0.0 Power = 1"
## [1] "Simple Effect Comparing M = 0.85 and 0.6. Power = 0.364"
## [1] "Simple Effect Comparing M = 0.85 and 0.85. Power = 0.05"
## [1] "Simple Effect Comparing M = 0 and 0.6. Power = 0.974"
```

$$\eta^2_{partial} = \frac{SS_{effect}}{SS_{effect} + SS_{error}} = \frac{45.18}{3011.1} = 0.0150$$

$$f^2 = \frac{\eta^2_{partial}}{1 - \eta^2_{partial}} = \frac{0.0150}{1 - 0.0150} = 0.0152$$

$$\lambda = f^2 df_{error} = 0.0152 \times 1000 = 15.2$$

R Code for Simple Effects

This test uses a function called anova2x2_se. The function requires the following input, this input is identical to the anova2x2 function for omnibus tests.

> anova2x2_se(m1.1, m1.2, m2.1, m2.2, s1.1, s1.2, s2.1, s2.2, n1.1, n1.2, n2.1, n2.2, alpha)

Table 5.9 shows the code and output for simple effects. When running this analysis, the output includes all possible simple effects tests and the means corresponding to those tests.

Note that the power for the contrast of interest, comparing the two justifications for tiebreaker policies, is .97. The power for comparing the two justifications for recruitment policies is is .05. Since the means for that contrast were equal, power is equal to α.

Power for Multiple Effects[1]

Designs involving multiple factors address two forms of power. The first is power for a single effect (e.g., a main effect or an interaction). This is the sort of power examined in the current chapter. Another conceptualization of power involves the power for detecting all effects in a specific design. I term this Power(All). For example in a study with two factors designed to yield power for both main effects and the interaction as .50, power for single effects reflects three different estimates [Power(A), Power(B), and Power(AxB)]. Power for detecting all effects on the other hand, reflects how likely it is to reject all three null hypotheses in the same study. You might be tempted to think this power

would be .50 as well. This is not the case. As a thought exercise, consider flipping a coin three times. The probability of the coin coming up heads is .50 on each flip. This is analogous to Power(A), Power(B), and Power(AxB) with each set at .50. However, the probability that the coin comes up heads on all three flips is far less than .50. Using the binomial approximation approach, this probability would be .13. This is analogous to Power(All) or how likely any one study is to reject all three null hypotheses.

Table 5.10 reflects Power(All) corresponding to various levels of power for individual effects. This table is simplistic as it only examines situations when the same level of power exists for each effect. In addition, this assumes that the factors are independent (i.e., participants randomly assigned to levels of the factor). Power(Each Effect) refers to a situation where both main effects and the interaction have equal power. The row labeled *p*(One or more reject) is the probability that one or more of the effect are detected and *p*(Two or more reject) is the probability that two or more effects are rejected. The row labeled *p*(Reject all H_0) is the probability that all three null hypotheses are rejected. Note that to obtain Power = .80 for rejecting all three null hypotheses, we would need to design for each test to have Power = .93. Table 5.11 shows the same calculations for a three factor ANOVA. In this table, the probabilities for rejecting at least one H_0 and for rejecting all the H_0*s* become more extreme.

Another way to think about power for multiple effects is in terms of Type I and Type II error rates. This issue is similar to inflation of α or Type I error. When conducting multiple significance tests, Type I error rates for the family of

TABLE 5.10 Power for Rejecting All Effects (and At Least One) for Various Levels of Individual Effect Power for Two Factor ANOVA

Power (Each effect)	.50	.60	.70	.80	.90	.93	.95
p(One or more reject H_0)	.88	.94	.97	.99	>.99	>.99	>.99
p(Two or more reject H_0)	.50	.65	.78	.90	.97	.99	>.99
p(Reject all H_0)	.13	.22	.34	.51	.73	.80	.86

TABLE 5.11 Power for Rejecting All Effects (and At Least One) for Various Levels of Individual Effect Power for Three Factor ANOVA

Power (Each effect)	.50	.60	.70	.80	.85	.90	.95	.964
p(One or more reject H_0)	.98	>.99	>.99	>.99	>.99	>.99	>.99	>.99
p(Two or more reject H_0)	.89	.96	.99	>.99	>.99	>.99	>.99	>.99
p(Three or more reject H_0)	.66	.82	.93	.98	>.99	>.99	>.99	>.99
p(Four or more reject H_0)	.34	.54	.74	.90	.98	>.99	>.99	>.99
p(Five or more reject H_0)	.11	.23	.42	.66	.89	.97	.98	>.99
p(Reject all H_0)	.02	.05	.12	.26	.38	.53	.74	.80

tests (a.k.a., familywise alpha) increase. Equation 5.24 provides an estimate of familywise α error for multiple comparisons and is the conceptual basis for development of tests such as the Bonferroni adjustment. According to the formula, with three tests using a pairwise alpha (α_{pw}) of .05, familywise alpha (α_{fw}) is .14.

$$\begin{aligned}\alpha_{fw} &= 1-(1-\alpha_{pw})^c \\ \alpha_{fw} &= 1-(1-.05)^3 = .14\end{aligned} \tag{5.24}$$

The same process is at work with regard to the familywise probability of making a β or Type II error (Formula 5.25), a value referred as β_{fw} in the formula. For example, take a study designed for β of .20 (called β_{ind} for Beta individual) for each of its three effects (a.k.a., Power = .80 for both main effects and the interaction). The likelihood of making a single β error among those three tests is substantially higher than the error rate of .20 for the individual tests. Just as with α error, multiple tests inflate the chances to make a single β error among a set of significance tests. The β_{fw} value easily converts to power to detect all of the effects in the design by taking $1-\beta_{fw}$. For three factors in this example, $\beta_{fw} = .488$ and Power(All) = .512.

$$\begin{aligned}\beta_{fw} &= 1-(1-\beta_{ind})^c \\ \beta_{fw} &= 1-(1-.20)^3 = .488\end{aligned} \tag{5.25}$$

For ANOVA designs where assignment to factor levels is random, calculation of Power(All) is straightforward (as the factors are not correlated). Simply multiple the power of the main effects and interaction together to obtain the estimate. In many ANOVA designs, Power(All) may not be relevant as it is common for research using such designs to only be interested in the interaction term. However, as discussed in Chapter 9, this issue becomes more complicated when dealing with designs that involve correlated predictors (e.g., multiple regression).

Table 5.12 provides an example that obtains Power(All) for the analyses in the present example. The anova2x2 function simply adds all = "ON" to produce the analysis.

TABLE 5.12 R Code and Output for Power(All)

```
anova2x2(m1.1=0.85, m1.2=0.85, m2.1=0.00, m2.2=0.60,
s1.1=1.7, s1.2=1.7, s2.1=1.7, s2.2=1.7,
n1.1=100, n1.2=100, n2.1=100, n2.2=100,
alpha=.05, all="ON")
## [1] "Power for Main Effect Factor A = 0.898"
## [1] "Power for Main Effect Factor B = 0.421"
## [1] "Power for Interaction AB = 0.421"
## [1] "Power(All) = 0.159"
```

Additional Issues

Additional issues focus on artificial dichotomization of continuously scaled predictors.

Artificial Dichotomization

An approach sometimes used with ANOVA designs involves taking continuously scaled variables and dichotomizing those values to create a factor. For example, scores on a self-esteem scale might be collected for a sample, then participants classified as high or low in self-esteem based on a median split of scores to create two roughly equal groups.

The best advice regarding this approach is do not dichotomize. Use regression instead. Artificially dichotomizing variables reduces power (see Cohen, 1984; Fitzsimons, 2008). Regression analysis is a more complicated statistical approach but regression yields more power as the variable remains in its original continuously scaled format.

Other important reasons for avoiding dichotomization also deserve mention. It is possible that dichotomizing produces groups that do not reflect clearly differentiated categories. For example, analyses of self-esteem data indicated a tendency to dichotomize self-esteem scale scores into high and low self-esteem categories based on median scores (Aberson, Healy, & Romero, 2000). Much of this dichotomization resulted in questionable classification of individuals as having "low self-esteem." Many individuals who scored moderately high on the scale end up classified as low self-esteem. For data on several major scales participants with scores that reflected 70% of the total possible score (e.g., a score of 70 on scale ranging from 1 to 100) were classified as "low self-esteem" despite scores that would more accurately be termed "medium self-esteem."

Dichotomization influences power through reduction of observed effect size. Formula 5.26 presents the attenuation factor statistic (Hunter & Schmidt, 1990) and Formula 5.27 presents the effect size attenuation statistic. In Formulae 5.26, a_d refers to attenuation due to dichotomization. The value $\Phi(c)$ is the unit normal density function for the z-transformed cutpoint (i.e., the "height" of the normal distribution curve). To command used to compute this value with R is dnorm(0,0,1). The first 0 reflects the cutpoint. The 0 and 1 that follow are the mean and standard deviation (use 0 and 1 to make this a normal distribution calculation). The values p and q reflect the proportion of participants in each group. If using a median split approach, $p = .50$ and $q = .50$, producing $\Phi(c) = .40$.

$$a_d = \frac{(c)}{\sqrt{pq}} \tag{5.26}$$

Equation 5.27 demonstrates the influence of artificial dichotomization on the effect size. The attenuating factor (a_d) reduces the size of the observed effect.

$$d_{observed} = d_{actual} a_d \tag{5.27}$$

The example that follows reflects a typical study utilizing dichotomization based on a median split. If population effect size is 0.50 and $a_d = .80$, the observed effect sizes is 0.40.

$$a_d = \frac{0.40}{\sqrt{0.5 * 0.5}} = \frac{0.40}{0.50} = 0.80$$
$$d_{observed} = 0.50 * 0.80 = 0.40$$

Regarding the influence of artificial dichotomization on power, a study designed to detect effects of $d = 0.50$ for a two-group design would require a sample size of $n = 128$ for Power = .80. However, the observed effect size does not accurately reflect the population effect. In this case, the observed effect size, $d_{observed} = 0.40$, with a sample of $n = 128$, yields Power = .61. If you must dichotomize, then recognize the influence this has on effect sizes and adjust sample sizes accordingly.

Summary

This chapter presented tests for one and two factor between subjects designs. These designs require estimation of meaningful patterns of means and accurate standard deviations. One factor designs require means across levels of each factor whereas two factor designs require cell means. A primary issue with both designs is whether hypotheses reflect omnibus tests or specific comparisons (e.g., planned contrast, simple effects). Well-developed hypotheses often predict outcomes best addressed through specific comparisons rather than omnibus tests. Power for specific comparisons often differs considerably from omnibus power.

Note

1. This section relies heavily on work by Maxwell (2004). I urge interested readers to consult this article as it details aspects of this issue that the present chapter does not address.

6

WITHIN SUBJECTS DESIGNS WITH ANOVA AND LINEAR MIXED MODELS

Within subjects (also known as repeated measures) designs focus on approaches involving measurement of the same participants at multiple levels of a factor (also known as independent variable). Often these designs involve measurement over two or more time periods. This chapter examines power for one and two factor within subjects Analysis of Variance (ANOVA) designs and trend analyses. Examples focus on calculations and analyses using univariate and linear mixed model (LMM) approaches, and present R functions for primary analyses, sphericity-adjusted tests, and trends.

Necessary Information

As with between subjects ANOVA designs, a good starting point is determining meaningful patterns of means (μs) and estimates of standard deviation (σ) for each factor level or cell. Also necessary are the expected correlations (ρs) between dependent measures.

Factors Affecting Power

Larger effect sizes and stronger positive correlations between dependent measures yield more power. Conceptually, correlations between measures explain variance that is otherwise attributed to error. The reduction of error when employing repeated measures makes within subjects designs more powerful than between subjects designs. As with other designs increases in sample size, α, and decreases in standard deviation increase power.

Although within subjects designs have great advantages regarding power, they also present an additional challenge. Designs that include three or more

levels of the within subjects factor come with an additional test assumption, namely the sphericity assumption. Sphericity is a complex issue that is discussed nicely elsewhere (e.g., Field, 1998). A simple (but incomplete) way to understand sphericity is that the assumption is satisfied if the correlations between each pair of measures are similar and variances across measures are homogeneous. Measures taken close together usually show higher correlations than those taken further apart, so the sphericity assumption is often violated. Sphericity assumption violations increase Type I error rates. As Type I error rates rise, so does power. This increase in power is fleeting; adjustments exist to account for this violation and drive the Type I error (and power) down.

This chapter presents two strategies for addressing violations of the sphericity assumption. The first approach involves downward adjustment of degrees of freedom in the univariate tests to account for inflated Type I error rates. This strategy, commonly termed epsilon adjustment, employs procedures such as the Greenhouse–Geisser (G–G) or Huynh–Feldt (H–F) statistics. Another approach involves use of LMM. LMM does not assume sphericity, so adjustments are unnecessary.

In the first edition of this text, I included a section on doing repeated measures via Multivariate Analysis of Variance (MANOVA). MANOVA approaches also do not require the sphericity assumption and are generally a bit more powerful than univariate ANOVA when assumptions are violated and sample sizes exceed two cases per dv (Tabachnick & Fidell, 2007b). However, the behavioral sciences appear to be shifting away from use of both ANOVA and MANOVA approaches for repeated measures and increasingly embracing LMM.

The main advantage of LMM over both ANOVA (and MANOVA approaches) is that LMM does not require complete data for participants. For example, a study with measures at three time points often includes participants that miss one or more of the measurement periods. ANOVA/MANOVA excludes participants with any missing data. LMM includes those participants with incomplete data (sometimes referred to as accommodating unbalanced designs). Practically, this means that LMM retains participants that missed any measurement period whereas ANOVA/MANOVA throws them out (i.e., uses listwise deletion). Including these participants affords LMM more power than ANOVA/MANOVA in these situations. This advantage does not influence the process of power analysis, but it does provide major benefits after data collection.

The present chapter addresses both univariate tests with epsilon adjustments and LMM approaches to repeated measures. Given the advantages LMM shows in dealing with missing data, my recommendation is to use LMM whenever possible.

Key Statistics

Univariate ANOVA

This chapter does not include discussion of sums of squares and related statistics. These calculations are more involved than those for between subjects ANOVA. For this reason, the techniques presented rely on computer-generated calculations for most values. Several sources provide excellent overviews of these calculations (e.g., Keppel, 1991).

Formulae 6.1 and 6.2 present two measures of effect size, partial η^2 and f^2. Partial η^2 is the more commonly presented effect size for ANOVA designs, reflecting the proportion of variance explained by a factor while partialing out the effects of other factors. In within subjects ANOVA, there are several error terms, so take care to choose the right one for this calculation (more on this in the calculation example). The value f^2 is less commonly presented in conjunction with significance test statistics, but it is necessary for calculating the noncentrality parameter (NCP). The NCP (Formula 6.3) is a function of the size of the effect size (partial η^2) and df_{error}.

$$\eta^2 = \frac{SS_{Effect}}{SS_{Effect} + SS_{Error}} \tag{6.1}$$

$$f^2 = \frac{\eta^2}{1-\eta^2} \tag{6.2}$$

$$\lambda = f^2 df_{error} \tag{6.3}$$

The formulae presented in this text differ slightly from values used in other sources (e.g., G*Power3). With regard to effect size, some programs request partial η^2 but define this value using formulae that do not account for correlations among dependent measures (these approaches adjust for the correlations later). The value presented in this chapter is consistent with what most sources and statistical software calls partial η^2. Specifically, the variance explained by other variables in the model are partialed out (e.g., in a two-factor design, the partial η^2 for factor A is SS_A/SS_A+SS_{Error}). Similarly, some sources calculate λ based on sample size and levels of the factor rather than degrees of freedom for error. I prefer to use degrees of freedom as it is more conservative (df_{error} is always less than the sample size). Different power estimates produced through different approaches likely reflect this choice.

Linear Mixed Models (LMM)

The NCP in LMM is the likelihood ratio Chi-square value derived by comparing the fit of a null model with a model that includes the predictor(s) of interest. This is often referred to as –2log or –2LL (see Formula 6.4).

$$\lambda = -2(ll_{Model} - ll_{Null}) \tag{6.4}$$

Sphericity Adjustments

Sphericity adjustment for univariate tests require a value called epsilon (ε). The ε value ranges from 0 to 1.0 with 1.0 meaning no violation of the assumption. Multiplying ε by degrees of freedom (*df)* produces adjusted df that make it more difficult to reject the null hypotheses. This approach is analogous to the *df* adjustment seen in Chapter 3 for *t*-tests with unequal variances. An alternative to using ε adjusted values, is to use LMM which does not assume sphericity.

Example 6.1: One Factor Within Subjects Design

The example for power for a one factor within subjects design focuses on a project designed to modify implicit attitudes through stereotype negation training. One study involved measures of implicit attitudes toward gay men taken at pretest, posttest (after training), 2 hours after posttest, and 6 hours after posttest. In between the pretest and posttest, participants engaged in a stereotype negation task that forced them to categorize nonstereotypical words with pictures of gay and heterosexual couples.

Determining the size of a meaningful effect required judgments regarding what size effect would be worth the time and effort required to develop and administer the test. Other researchers had success with similar approaches in improving attitudes toward other groups, so we were interested in detecting similar size outcomes. That work found raw score changes of +0.25 to +0.40 (meaning more positive attitudes) from pre to post and gradual increases thereafter. Based on this information, we judged +0.25 as the minimum value for a practically important pre-post change. That is, to term the technique effective, it was important to achieve at worst the same level of attitude change as the least-effective previous study.

We were also interested in whether attitudes continued to change over time and judged smaller increases as meaningful at 2 and 6 hours. Estimates of standard deviations relied on reported uses of similar dependent measures, yielding an estimate of $\sigma = 0.40$ with slight increases in variability for each subsequent measure. For the correlations between measures, large-scale studies of attitudes toward other groups reported test–retest correlation of .50. However, correlations tend to degrade over time, and the test–retest reliabilities for similar measures expressed a considerable range across studies, suggesting smaller correlations for measures further apart (e.g., correlations for pre-twohour = = .30, correlations for pre-sixhour = .15).

Table 6.1 presents the estimates of population means, variances, and correlations needed for establishing power. Many of the values reflect conservative estimates. In general, smaller correlations between measures mean less power, larger standard deviations mean less power, and greater heterogeneity of variances means less power. Also, the divergent correlation values promote violation of the sphericity assumption.

TABLE 6.1 Descriptive Statistics for Within Subjects ANOVA Example

	Pre	*Post*	*2 Hour*	*6 Hour*
Pre	$\mu = -0.25$ $\sigma = 0.40$	–	–	–
Post	$\rho = .50$	$\mu = 0.00$ $\sigma = 0.50$	–	–
Two Hour	$\rho = .30$	$\rho = .50$	$\mu = 0.10$ $\sigma = 0.60$	–
Six Hour	$\rho = .15$	$\rho = .30$	$\rho = .50$	$\mu = 0.15$ $\sigma = 0.70$

The code and output in Table 6.2 demonstrate use of the win1F function. The format of the function is:

win1F(m1, m2, m3, m4, s1, s2, s3, s4, r12, r13, r14, r23, r24, r34, n, alpha)

The values m1-m4 and s1-s4 reflect means and standard deviations for each factor level. The *r* values correspond to correlations between the dvs. The code allows for two to four factors. Leave out values not relevant to your analyses. For example, if you have three factors, omit, m4, s4, r14, r24, and r34. The value *n* is overall sample size. Alpha defaults to .05 if no value is entered.

Based on the univariate tests (also known as univariate unadjusted or sphericity assumed) in Table 6.2, a sample of 25 participants yields power of .809.

Table 6.3 provides output from an ANOVA using the means, standard deviations, correlations, and sample sizes from the present analysis. This is presented to demonstrate where the various values for calculation come from, it is not a step to determine power. Regarding calculations, η^2 uses $SS_{error} = 14.53$ and $SS_{effect} = 2.38$ (bolded in Table 6.3). The noncentrality parameter (NCP) (λ) uses the value labeled DFd for iv = 72. Computations use Formulae 6.1, 6.2, and 6.3.

TABLE 6.2 R Code and Output for One Factor Within Design using ANOVA

```
win1F(m1=-.25, m2=.00, m3=.10, m4=.15, s1=.4, s2=.5, s3=.6, s4=.7,
r12=.50, r13=.30, r14=.15, r23=.5, r24=.30, r34=.50, n=25)
## [1] "Power (Unadjusted) for n = 25 = 0.809"
## [1] "Power H-F Adjusted (Epsilon = 0.914) for n = 25 = 0.782"
## [1] "Power G-G Adjusted (Epsilon = 0.815) for n = 25 = 0.745"
```

TABLE 6.3 Information for One Factor ANOVA Calculation Example

```
## $ANOVA
##  Effect DFn DFd SSn   SSd        F          p p<.05        ges
## 1 (Intercept) 1 24 0.000 15.708 0.000000 1.00000    0.00000000
## 2          iv 3 72 2.375 14.532 3.922378 0.01187129 0.07281925
```

$$\eta^2 = \frac{SS_{Effect}}{SS_{Effect} + SS_{Error}} = \frac{2.38}{2.38 + 14.53} = 0.141$$

$$f^2 = \frac{\eta^2}{1 - \eta^2} = \frac{0.141}{1 - 0.141} = 0.164$$

$$\lambda = f^2 df_{error} = 0.164 \times 72 = 11.8$$

Example 6.2: Sphericity Adjustments

The output in Table 6.2 also displays the epsilon adjustments (Greenhouse–Geisser etc.). Often data collected using within subjects designs do have a sphericity problem, so I recommend a conservative approach that assumes a problem exists. Satisfying the conservative assumption yields adequate power when violating assumptions and even better power if there is no assumption violation. The output shows Greenhouse–Geisser $\varepsilon = .815$. Although there are some situations where Huynh–Feldt is the preferred statistic for epsilon adjustment, I recommend G–G, for use in power analysis because is the more conservative approach. The sphericity adjustment is applied to the *df*. For example, the *df* involved in the significance test for Time are 3 and 72. The G–G adjustment multiplies the G–G epsilon value of .815 by each *df*, producing 2.44 (3 * .815) and 58.65 (72 * .815). These values become the degrees of freedom for that adjusted test. The *df*s based on these adjustments are then used along with the *F* statistic to calculate the probability [$F(2.44, 58.65) = 3.92$, $p = .018$]. For researchers who prefer the H–F adjustment, the approach detailed in Table 6.2 works in the same manner, simply substitute ε of .914 for .815. Regardless of the approach, both adjustments suggest a slightly larger sample to obtain power of .80.

Example 6.3: Linear Mixed Model Approach to Repeated Measures

LMM is an increasingly popular alternative to the univariate within subjects ANOVA approaches demonstrated earlier. As noted earlier in this chapter, LMM has two major advantage: it does not require the sphericity assumption and it retains participants with missing data. The example that follows first presents output from a LMM analysis (shown in Table 6.4 and corresponding to the values in Example 6.2).

$$\lambda = -2(-80.22 - -74.54) = 11.35$$

TABLE 6.4 Information for One Factor LMM Calculation Example

```
##    Model df      AIC      BIC  logLik    Test  L.Ratio p-value
## base   1  4 168.4380 178.8587 -80.21902
## model1 2  7 163.0849 181.3211 -74.54247 1 vs 2 11.3531   0.01
```

TABLE 6.5 R Code and Output for One Factor Within Design using LMM

```
lmm1F(m1=-.25, m2=.00, m3=.10, m4=.15, s1=.4, s2=.5, s3=.6, s4=.7,
r12=.50, r13=.30, r14=.15, r23=.5, r24=.30, r34=.50, n=25)
## [1] "Power for n = 25 = 0.817"
```

Using the code 1-pchisq(7.815, 3, 11.35) where 7.815 represents Chi square at .05 for 3 degrees of freedom (the second value in the code), yields power of .82 for the LMM test.

Table 6.5 presents code and output for the LMM approach using the lmm1F function. The values required by the function are as follows and are the same as those values for win1F demonstrated in Example 6.2.

lmm1F(m1, m2, m3, m4, s1, s2, s3,s4, r12, r13, r14, r23, r24, r34, n, alpha)

Example 6.4: A Serious Sphericity Problem

To demonstrate how power for the univariate adjusted and LMM tests differ when violating the sphericity assumption, I modified Example 6.1 to represent worse problems with sphericity. The code in Table 6.6 demonstrates issues that contribute to violation of the sphericity assumption. When compared to the parameters in Table 6.1, the standard deviations are more unequal and the correlations between measures more divergent. For this analysis, with a sample of 100, the G–G statistic is smaller than in the previous example (.662 vs. .815), reflecting a greater deviation from the sphericity assumption. Under these conditions, LMM provides a bit more power than the adjusted tests.

Trend Analysis

Often research involving measures taken over time seek to examine trends. For example, a linear trend might exist where scores rise (or fall) over time or a quadratic (curvilinear) trend shows that scores improve initially but later return to pretest levels. This is a qualitatively different question than the one addressed in the example where power analysis focused on tests examining whether the four means (one for each time) differed. Trend analyses ask whether means

TABLE 6.6 Code and Output for Serious Sphericity Problem Example

```
win1F(m1=-.25, m2=.00, m3=.10, m4=.15, s1=.4, s2=.5, s3=2.5, s4=2.0,
r12=.50, r13=.30, r14=.10, r23=.5, r24=.30, r34=.40, n=100)
## [1] "Power (Unadjusted) for n = 100 = 0.397"
## [1] "Power H-F Adjusted (Epsilon = 0.675) for n = 100 = 0.321"
## [1] "Power G-G Adjusted (Epsilon = 0.662) for n = 100 = 0.318"
lmm1F(m1=-.25, m2=.00, m3=.10, m4=.15, s1=.4, s2=.5, s3=2.5, s4=2.0,
r12=.50, r13=.30, r14=.10, r23=.5, r24=.30, r34=.40, n=100)
## [1] "Power for n = 100 = 0.403"
```

differ in a specific manner. Trend analyses do not require a sphericity adjustment. Trends involve a single degree of freedom in the numerator. Sphericity is not an issue when $df_{num} = 1$.

Example 6.5: Trend Analysis

Imagine that in the example from the previous section, hypotheses focused on steady and consistent changes in attitudes over each measurement period. For example, instead of predicting different levels of improvement across each measure (e.g., +0.25, then +0.10, then +0.05), we expected a +0.10 improvement in attitudes for each measurement period. This prediction reflects a hypothesis about a linear trend rather than an omnibus ANOVA test that simply specifies differences between means. This distinction is important as the two types of tests (omnibus vs. trend analysis) sometimes produce markedly different power analyses.

The R code and output in Table 6.7 modified Example 6.1 to represent 0.10 mean improvements over at each time period using both the ANOVA and LMM approaches. The format of the functions that follows is the same as those in previous analyses, the only difference is the name of the functions.

win1Ftrends(m1, m2, m3, m4, s1, s2, s3, s4, r12, r13, r14, r23, r24, r34, n, alpha)

lmm1Ftrends(m1, m2, m3, m4, s1, s2, s3, s4, r12, r13, r14, r23, r24, r34, n, alpha)

The output contains two tests for each effect for the ANOVA test. Researchers differ in preference for degrees of freedom with some using the larger value (72) taken from the denominator of the omnibus ANOVA and others splitting the *df* between tests (24).

Notice that the power for detecting the linear trend (.69) for the LMM is considerably higher than the power for the ANOVA (.49) with $n = 25$. The

TABLE 6.7 R Code and Output for Trend Analysis

```
win1Ftrends(m1=-.25, m2=-.15, m3=-.05, m4=.05, s1=.4, s2=.5,
s3=.6, s4=.7,
r12=.50, r13=.30, r14=.15, r23=.5, r24=.30, r34=.50, n=25)
## [1] "Power Linear Trend for n = 25 df = 24 = 0.487"
## [1] "Power Linear Trend for n = 25 df = 72 = 0.508"
## [1] "Power Quadratic Trend for n = 25 df = 24 = 0.05"
## [1] "Power Quadratic Trend for n = 25 df = 72 = 0.05"
## [1] "Power Cubic Trend for n = 25 df = 24 = 0.05"
## [1] "Power Cubic Trend for n = 25 df = 72 = 0.05"
lmm1Ftrends(m1=-.25, m2=-.15, m3=-.05, m4=.05, s1=.4, s2=.5,
s3=.6, s4=.7,
r12=.50, r13=.30, r14=.15, r23=.5, r24=.30, r34=.50, n=25)
## [1] "Power Linear Trend for n = 25 = 0.69"
## [1] "Power Quadratic Trend for n = 25 = 0.05"
## [1] "Power Cubic Trend for n = 25 = 0.055"
## [1] "Tests use df = 72"
```

pattern of means contains no quadratic or cubic component, so power for these trends are low for both analyses.

Power for the specific contrasts or patterns of effect of interest often produce different power than omnibus tests. The best question to ask is what the effect of interest is. Does the research hypothesis focus on the trend, the omnibus test, or both? In addition, what trend is important? Whatever your interest, design for detecting that effect.

Example 6.6: Two Within Subject Factors Using ANOVA

The previous examples examined a single within subjects factor. The example that follows adds a second within subjects factor (length of negation task). Participants complete a short negation task (15 minutes; Condition A) then return a month later and complete a longer negation task (1 hour; Condition B). Both conditions involve pre, post, 2 hour, and 6-hour measures.

For simplicity, the code reflects all the correlations as .5. The code does accept a full correlation matrix. For this example that would require correlation estimates for all eight measurement periods. Unless you can provide very good estimates of those values, it is best to choose a conservative overall estimate of the correlation. The means reflect the expectation that the longer test produces stronger attitude change. The difference between the means for the short and long tasks reflect the size of effect that justifies increasing the length of the intervention. For this design, there are two main effects and an interaction.

The code and output in Table 6.8 demonstrate use of the win2F function. The format of the function is:

```
win2F(m1.1, m2.1, m3.1, m4.1, m1.2, m2.2, m3.2, m4.2, r, s1.1, s2.1, s3.1,
s4.1, s1.2, s2.2, s3.2, s4.2, n, alpha)
```

The values m1.1 through m4.2 reflect cell means and s1.1 through s4.2 reflect cell standard deviations. The notation places the level of the first factor before the decimal and the level of the second factor after the decimal. For equal standard deviations, provide one value for *s* (e.g., s = 1.5). The value *r* reflects correlations between dvs (use r12, r13, etc. to specify individual correlations). The code allows for two to four levels on the first factor and two levels on the second factor. Leave out values not relevant to your analyses. For example, if you have three factors, omit, m4.1 and m4.2. The value *n* is overall sample size. Alpha defaults to .05 if no value is entered.

The analysis in Table 6.8 with $n = 80$ demonstrates good power for Time (Factor A) but lackluster power for Condition (Factor B) and the interaction. If the research goal is to determine that one condition outperforms another, then a sample of 80 participants is too small to ensure adequate power.

The second analysis in the table finds sample size for a design that yields power of at least .80 for both main effects. This analysis used a sample of $n = 337$.

TABLE 6.8 R Code and Output for Two Factor Within Design using ANOVA

```
win2F(m1.1=-.25, m2.1=0, m3.1=.10, m4.1=.15, m1.2=-.25,
m2.2=.10, m3.2=.30, m4.2=.35, s1.1=.4, s2.1=.5, s3.1=2.5,
s4.1=2.0, s1.2=.4, s2.2=.5, s3.2=2.5, s4.2=2.0, r=.5, n=80)
## [1] "Power Factor A (Unadjusted) for n = 80 = 0.748"
## [1] "Power Factor A H-F Adjusted (Epsilon = 0.61) for
n = 80 = 0.592"
## [1] "Power Factor A G-G Adjusted (Epsilon = 0.597) for
n = 80 = 0.585"
## [1] "Power Factor B (Unadjusted) for n = 80 = 0.272"
## [1] "Power Factor B Adjusted - There is no adjustment when
levels = 2"
## [1] "Power Factor AB (Unadjusted) for n = 80 = 0.102"
## [1] "Power Factor AB H-F Adjusted (Epsilon = 0.628) for
n = 80 = 0.092"
## [1] "Power Factor AB G-G Adjusted (Epsilon = 0.614) for
n = 80 = 0.091"
win2F(m1.1=-.25, m2.1=0, m3.1=.10, m4.1=.15, m1.2=-.25,
m2.2=.10, m3.2=.30, m4.2=.35, s1.1=.4, s2.1=.5, s3.1=2.5,
s4.1=2.0, s1.2=.4, s2.2=.5, s3.2=2.5, s4.2=2.0, r=.5, n=337)
## [1] "Power Factor A (Unadjusted) for n = 337 = 1"
## [1] "Power Factor A H-F Adjusted (Epsilon = 0.6) for
n = 337 = 0.997"
## [1] "Power Factor A G-G Adjusted (Epsilon = 0.597) for
n = 337 = 0.997"
## [1] "Power Factor B (Unadjusted) for n = 337 = 0.8"
## [1] "Power Factor B Adjusted - There is no adjustment when
levels = 2"
## [1] "Power Factor AB (Unadjusted) for n = 337 = 0.314"
## [1] "Power Factor AB H-F Adjusted (Epsilon = 0.618) for
n = 337 = 0.246"
## [1] "Power Factor AB G-G Adjusted (Epsilon = 0.614) for
n = 337 = 0.245"
```

Example 6.7: Simple Effects Using ANOVA

As with between subjects factorial ANOVA, when interactions exist, designs often probe interactions using simple effects tests. For example, we might be interested in showing significant improvements in attitudes for both groups individually. That is, that both techniques did in fact change attitudes for the better.

The code and output in Table 6.9 demonstrate use of the win2Fse function. The format of the function is the same as for win2f:

> win2Fse(m1.1, m2.1, m3.1, m4.1, m1.2, m2.2, m3.2, m4.2, r, s1.1, s2.1, s3.1, s4.1, s1.2, s2.2, s3.2, s4.2, n, alpha)

Table 6.9 shows the code and output for these analyses using the $n = 337$ sample size. Output includes all possible simple effects. Based on this analysis, there

TABLE 6.9 R Code and Output for Two Factor Within Design With Simple Effects Using ANOVA

```
win2Fse(m1.1=-.25, m2.1=0, m3.1=.10, m4.1=.15, m1.2=-.25,
m2.2=.10, m3.2=.30, m4.2=.35, s1.1=.4, s2.1=.5, s3.1=2.5,
s4.1=2.0, s1.2=.4, s2.2=.5, s3.2=2.5, s4.2=2.0, r=.5, n=220)
## [1] "Power Factor A at B1 (Unadjusted) for n = 220 = 0.803"
## [1] "Power Factor A at B1 H-F Adjusted (Epsilon = 0.652)
for n = 220 = 0.668"
## [1] "Power Factor A at B1 G-G Adjusted (Epsilon = 0.647)
for n = 220 = 0.666"
## [1] "Power Factor A at B2 (Unadjusted) for n = 220 = 0.994"
## [1] "Power Factor A at B2 H-F Adjusted (Epsilon = 0.652)
for n = 220 = 0.965"
## [1] "Power Factor A at B2 G-G Adjusted (Epsilon = 0.647)
for n = 220 = 0.964"
## [1] "Power Factor B at A1 for n = 220 = 0.05"
## [1] "Power Factor B at A2 for n = 220 = 0.84"
## [1] "Power Factor B at A3 for n = 220 = 0.219"
## [1] "Power Factor B at A4 for n = 220 = 0.315"
```

exists considerable power to detect differences in time across each condition (A at B1, A at B2). The analyses noted B at A1, etc. reflect comparisons across conditions at each of the four time periods.

Example 6.8: Two Factor Within and Simple Effects Using LMM

The code and output in Tables 6.10 and 6.11 provide power using LMM approaches. The format of the functions is identical to the win2F and win2Fse functions in the previous examples (except for the name of the function):

lmm2F(m1.1, m2.1, m3.1, m4.1, m1.2, m2.2, m3.2, m4.2, r, s1.1, s2.1, s3.1, s4.1, s1.2, s2.2, s3.2, s4.2, n, alpha)

TABLE 6.10 R Code and Output for Two Factor Within Design using LMM

```
lmm2F(m1.1=-.25, m2.1=0, m3.1=.10, m4.1=.15, m1.2=-.25,
m2.2=.10, m3.2=.30, m4.2=.35, s1.1=.4, s2.1=.5, s3.1=2.5,
s4.1=2.0, s1.2=.4, s2.2=.5, s3.2=2.5, s4.2=2.0, r=.5, n=80)
## [1] "Power Factor A for n = 80 = 0.752"
## [1] "Power Factor B for n = 80 = 0.28"
## [1] "Power AxB for n = 80 = 0.104"
lmm2F(m1.1=-.25, m2.1=0, m3.1=.10, m4.1=.15, m1.2=-.25,
m2.2=.10, m3.2=.30, m4.2=.35, s1.1=.4, s2.1=.5, s3.1=2.5,
s4.1=2.0, s1.2=.4, s2.2=.5, s3.2=2.5, s4.2=2.0, r=.5, n=337)
## [1] "Power Factor A for n = 337 = 1"
## [1] "Power Factor B for n = 337 = 0.802"
## [1] "Power AxB for n = 337 = 0.315"
```

TABLE 6.11 R Code and Output for Two Factor Within Design With Simple Effects Using LMM

```
lmm2Fse(m1.1=-.25, m2.1=0, m3.1=.10, m4.1=.15, m1.2=-.25,
m2.2=.10, m3.2=.30, m4.2=.35, s1.1=.4, s2.1=.5, s3.1=2.5,
s4.1=2.0, s1.2=.4, s2.2=.5, s3.2=2.5, s4.2=2.0, r=.5, n=220)
## [1] "Power A at B1 for n = 220 = 0.804"
## [1] "Power A at B2 for n = 220 = 0.994"
## [1] "Power B at A1 for n = 220 = 0.05"
## [1] "Power B at A2 for n = 220 = 0.837"
## [1] "Power B at A3 for n = 220 = 0.221"
## [1] "Power B at A4 for n = 220 = 0.317"
```

lmm2Fse(m1.1, m2.1, m3.1, m4.1, m1.2, m2.2, m3.2, m4.2, r, s1.1, s2.1, s3.1, s4.1, s1.2, s2.2, s3.2, s4.2, n, alpha)

Additional Issues

Issues related to detecting power for multiple effects, as discussed in Chapter 5, also pertain to within subjects designs. Recall that the power to detect all the effects of interest is a function [termed Power(All)] of the product of the power of all the tests. In the example with $n = 337$, Power(All) reflects the product of the three power values (1 * .8 * .34 = .27). Designing to find significance for all three effects in the same study requires larger sample sizes.

Summary

This chapter examined one and two factor within subjects designs. The primary information required for each design are meaningful patterns of means, standard deviations for each dependent measure, and correlations between measures. For standard deviations and correlations, accurate estimates improve power analysis. When expecting heterogeneous standard deviations or different correlations across measures (i.e., violation of the sphericity assumption), power analysis should address these issues through consideration of sphericity-adjusted approaches or use of LMM. Power analysis results for hypotheses specifying trends or simple effects often diverge from results for omnibus tests so power analysis should focus on the specific tests of interest.

7

MIXED MODEL ANOVA AND MULTIVARIATE ANOVA

This chapter presents power analyses for ANOVA or Linear Mixed Models (LMM) with both between and within subjects factors and Multivariate ANOVA (MANOVA). MANOVA address designs with multiple dependent measures and at least one factor. This chapter examines power for models with one between and one within subjects factor, one factor MANOVA, and discusses how different patterns of effect sizes and correlations in MANOVA influence power.

Necessary Information

As with the ANOVA designs in Chapters 5 and 6, power analysis requires means (μs) corresponding to meaningful differences (or patterns of differences) among factor levels, estimates of the standard deviation (σ) for each factor level, and the expected correlations (ρs) between dependent measures.

Factors Affecting Power

For ANOVA with between and within factors, larger effect sizes and stronger positive correlations between dependent measures yield more power. Sphericity influences power for within subjects factors as discussed in Chapter 6. As with other designs larger sample sizes and alpha increase power.

For MANOVA, the pattern of correlations between variables and the pattern of differences between means (e.g., effect sizes for each dependent measure) influence power. This is a complex issue, so I devote a chapter section to it.

Key Statistics

There are no new statistics introduced in this chapter. Calculation of effect sizes and noncentrality parameters (NCP) use Formulae 6.1–6.3, presented in detail

in Chapter 6. A full review of the statistics associated with these procedures and their calculation is outside the scope of this chapter. For a highly readable overview of the techniques, see Tabachnick and Fidell (2007a; 2007b).

ANOVA with Between and Within Subject Factors

This section examines power for ANOVA designs with at least one within subjects and one between subjects factor. Some sources refer to designs with both between and within factors as mixed model ANOVA, mixed randomized-repeated, or split-plot designs. As the term "mixed" is increasingly used to address approaches random and fixed factors (e.g., LMM), I refer to these designs as ANOVA with both between and within subjects factors.

Example 7.1: ANOVA with One Within Subjects Factor and One Between Subjects Factor

The example in this section expands the one factor within subjects study from Chapter 6 through addition of a between subjects factor. In the initial example, participants engaged in a stereotype negation procedure and our interest was whether attitudes improved over time. Another reasonable question could be whether this level of change differed from the change in the control group. To address this issue, the study in this example includes a control group that completed a stereotype maintenance procedure (e.g., Kawakami, Dovidio, Moll, Herrasen, & Russin, 2000). Participants in this condition engaged in a task that forced them to respond in stereotype-consistent rather than stereotype-negating manners. Earlier work demonstrated that participants who completed a maintenance task showed consistent attitudes across all levels of measurement. Therefore, in this example, we did not expect participants in the maintenance condition to experience changes in their attitudes.

Table 7.1 details the means, standard deviations, and correlations for the measures. Note that the expected control group means (μ_c) remain constant while the treatment group means (μ_t) change in the same manner as in the Chapter 6 example.

Table 7.2 presents the R code and output for this analysis. The code addresses correlations by group. The negation group is listed first and noted with a "1" (e.g., M1.1 = –.25, M2.1 = .00) and the maintenance groups appears second with a "2" (e.g., M1.2 = –.25, M2.2 = –.25).

The format of the function is as follows:

```
win1bg1(m1.1, m2.1, m3.1, m4.1, m1.2, m2.2, m3.2, m4.2,
s1.1, s2.1, s3.1, s4.1, s1.2, s2.2, s3.2, s4.2,
```

TABLE 7.1 Descriptive Statistics for Mixed Model ANOVA Example

	Pre	*Post*	*2 Hour*	*6 Hour*
Pre	$\mu_t = -0.25$, $\sigma_t = 0.40$ $\mu_c = -0.25$, $\sigma_c = 0.40$	–	–	–
Post	$\rho = .50$	$\mu_t = 0.0$, $\sigma_t = 0.50$ $\mu_c = -0.25$, $\sigma_c = 0.50$	–	–
Two Hour	$\rho = .30$	$\rho = .50$	$\mu_t = 0.10$, $\sigma_t = 0.60$ $\mu_c = -0.25$, $\sigma_c = 0.60$	–
Six Hour	$\rho = .15$	$\rho = .30$	$\rho = .50$	$\mu_t = 0.15$, $\sigma_t = 0.70$ $\mu_c = -0.25$, $\sigma_c = 0.70$

Note
Subscript *t* is the negation (treatment) group and subscript *c* is the maintenance (control) group.

TABLE 7.2 R Code and Output for ANOVA with One Between and One Within Factor Example

```
win1bg1(m1.1=-.25, m2.1=0, m3.1=0.10, m4.1=.15, m1.2=-.25,
m2.2=-.25,
m3.2=-.25, m4.2=-.25, s1.1=.4, s2.1=.5, s3.1=0.6, s4.1=.7,
s1.2=.4, s2.2=.5, s3.2=.6, s4.2=.7, n=50,
r1.2_1=.5, r1.3_1=.3, r1.4_1=.15, r2.3_1=.5, r2.4_1=.3,
r3.4_1=.5,
r1.2_2=.5, r1.3_2=.3, r1.4_2=.15, r2.3_2=.5, r2.4_2=.3,
r3.4_2=.5)
## [1] "Power Factor A (Between) for n = 50 = 0.864"
## [1] "Power Factor A H-F Adjusted (Epsilon = 0.837) for
n = 50 = 0.819"
## [1] "Power Factor A G-G Adjusted (Epsilon = 0.815) for
n = 50 = 0.812"
## [1] "Power Factor B (Within) for n = 50 = 0.827"
## [1] "Power Factor B Adjusted - There is no adjustment when
levels = 2"
## [1] "Power Factor AB (Unadjusted) for n = 50 = 0.827"
## [1] "Power Factor AB H-F Adjusted (Epsilon = 0.837) for
n = 50 = 0.761## [1] "Power Factor AB G-G Adjusted
(Epsilon = 0.815) for n = 50 = 0.765
```

r1.2_1, r1.3_1, r1.4_1, r2.3_1, r2.4_1, r3.4_1,

r1.2_2, r1.3_2, r1.4_2, r2.3_2, r2.4_2, r3.4_2,

r, s, n, alpha)

The values m1-m4 and s1-s4 reflect means and standard deviations for each factor level. These are defined by group with m1.1 meaning the mean at time 1 for the first group and m1.2 being the time 1 mean for the second group. The *r* values correspond to correlations between the dvs. The _1 or _2 defines the group. The code allows for two to four within subjects factors. Leave out values not relevant to your analyses. For example, if you have three factors, omit, m4.1, m4.2, s4.1, s4.2, r14, r24, and r34. The value *n* is overall sample size. Alpha defaults to .05 if no value is entered. The values *r* and *s* should be left blank if providing individuals means and standard deviations. For the same standard deviation or correlation across all values, omit the individual values (e.g., s1.1, r1.2_1) and enter a single value for *r* or *s*.

Example 7.2: Linear Mixed Model with One Within Subjects Factor and One Between Subjects Factor

As in Chapter 6, LMM approaches provide an alternative analysis approach. The function presented below performs the analysis from Example 7.1 using LMM. The format of the function is:

TABLE 7.3 R Code and Output for LMM with One Between and One Within Factor Example

```
lmm1w1b(m1.1=-.25, m2.1=0, m3.1=0.10, m4.1=.15, m1.2=-.25,
m2.2=-.25,
m3.2=-.25, m4.2=-.25, s1.1=.4, s2.1=.5, s3.1=0.6, s4.1=.7,
s1.2=.4, s2.2=.5, s3.2=.6, s4.2=.7, n=50,
r1.2_1=.5, r1.3_1=.3, r1.4_1=.15, r2.3_1=.5, r2.4_1=.3, r3.4_1=.5,
r1.2_2=.5, r1.3_2=.3, r1.4_2=.15, r2.3_2=.5, r2.4_2=.3, r3.4_2=.5)
## [1] "Power Factor A (Between) for n = 50 = 0.862"
## [1] "Power Factor B (Within) for n = 50 = 0.817"
## [1] "Power AxB for n = 50 = 0.833"
```

lmm1w1b(m1.1, m2.1, m3.1, m4.1, m1.2, m2.2, m3.2, m4.2,

s1.1, s2.1, s3.1, s4.1, s1.2, s2.2, s3.2, s4.2,

r1.2_1, r1.3_1, r1.4_1, r2.3_1, r2.4_1, r3.4_1,

r1.2_2, r1.3_2, r1.4_2, r2.3_2, r2.4_2, r3.4_2,

r, s, n, alpha)

The input values for the function do not differ from the win1bg1 function. Note that the first 1 in the function name is a lowercase l (the letter). The characters before w and b are 1 (the number one).

As shown in Table 7.3, the LMM approach provides slightly different power estimates. For each effect the LMM power is higher than power for the sphericity-adjusted tests.

Multivariate ANOVA

MANOVA procedures address comparisons across groups on two or more dependent variables (dv). For power analysis, MANOVA brings additional complexity. With MANOVA, patterns of effect sizes across the dependent measures and patterns of correlations between dependent measures influence power considerably. For that reason, discussion of how these patterns affect power appears before coverage of MANOVA power because understanding these relationships is important in designing for adequate power. As suggested throughout the text, when in doubt (and when it is feasible) it is good practice to design conservatively. An understanding of how the patterns of correlations and effects influence power helps to determine what is and what is not a conservative design decision.

Patterns of Effects and Correlations

Aside from the basic issues influencing power for all ANOVA designs, two new issues affect MANOVA power. The first involves the type of effect sizes

observed for each dependent measure. In general, if all the effect sizes are consistent, one pattern of power results exist whereas for inconsistent effects there is a different pattern. In this context, consistent means roughly the same effect size across all dependent measures. When effects are inconsistent (e.g., some dependent measures differ strongly across the between subjects factor and others show small differences), a different pattern of power results exists.

Power also depends on correlations between dependent measures. For consistent effects (e.g., Small–Small), power increases as we move from strongly positive to strongly negative correlations. For inconsistent effects (e.g., Small–Strong), power increases when moving toward more extreme relationships (either stronger positive or stronger negative correlations). Regardless of the relationship, negative correlations between predictors usually produce more power than positive correlations of the same magnitude (but see the discussion that follows regarding recoding variables).

Table 7.4 summarizes these relationships for a MANOVA with two dependent measures (see also Cole, Maxwell, Arvey, & Salas, 1994 for a technical description). All situations in the table use $n = 20$ except for the Small–Small effect column where tests used $n = 50$ to more clearly demonstrate the pattern of results. In the table, small refers to a $d = 0.20$ between two conditions on a single dv. Moderate refers to $d = 0.50$ and strong indicates a $d = 0.80$. The column labeled Small–Small reflects when two dv both show $d = 0.20$, in the Small–Strong column one measure has $d = 0.20$ and the other $d = 0.80$, and so on. All effects in the table represent results in the same direction (e.g., both positive).

One interpretation drawn from Table 7.4 is that negative correlations increase power. This leads to an obvious question. Every time I teach or present on this topic someone asks, "If negative correlations between dependent measures increase my power, does this mean I can reverse code one of my dvs to yield more power?" The answer is no. Table 7.4 presents power for situations where predictors relate to the dependent measure in the same manner (e.g., Group 1 scores higher than Group 2 on both measures). That is, the independent variable (iv)–dv relationships are all in the same direction. If one variable were reversed coded (e.g., high scores converted to low scores), a reversal of the direction of the effect size would follow for that variable. To obtain values for situations where one iv has a positive and one has a negative relationship with the dv, the values in the correlation table reverse. For example, with $d = 0.5$ for one iv and $d = -0.5$ for the other, and the dvs correlated at $-.40$, the power would be .34 (the value listed for $r = .40$) rather than .69 which is the value for $r = -.40$ (see the Additional Issues section for a more detailed explanation).

Table 7.4 illustrates several other important considerations for MANOVA. First, power tends to be higher when negative correlations exist between dependent measures. Practically if iv–dv relationships are in the same direction, strong negative correlations between dependent measures are not common.

TABLE 7.4 Power as a Function of Effect Size and Correlation Patterns

Correlation Between Measures	*Small–Small d = 0.2, 0.2*	*Moderate–Moderate d = 0.5, 0.5*	*Strong–Strong d = 0.8, 0.8*	*Small–Moderate d = 0.2, 0.5*	*Small–Strong d = 0.2, 0.8*	*Moderate–Strong d = 0.5, 0.8*
.9	.14	.26	.59	.52	.98	.72
.8	.14	.27	.61	.34	.84	.60
.7	.15	.29	.64	.29	.72	.57
.6	.15	.30	.67	.26	.64	.56
.5	.16	.32	.70	.25	.60	.57
.4	.17	.34	.73	.25	.58	.59
.3	.18	.36	.76	.25	.57	.61
.2	.19	.39	.80	.26	.57	.64
.1	.20	.42	.83	.27	.57	.68
.0	.22	.46	.87	.28	.59	.72
−.1	.24	.50	.90	.30	.62	.76
−.2	.26	.55	.93	.33	.65	.81
−.3	.30	.61	.96	.36	.70	.86
−.4	.34	.69	.98	.41	.75	.91
−.5	.40	.77	.99	.47	.82	.95
−.6	.49	.86	1.0	.56	.89	.98
−.7	.61	.94	1.0	.69	.96	1.0
−.8	.80	.99	1.0	.86	.99	1.0
−.9	.98	1.0	1.0	.99	1.0	1.0

When negative correlations are present, they are usually small. However, small negative correlations give more power than small positive correlations. Second, in several cases strong positive correlations between dependent measures reduce power. If strong positive correlations exist between dependent measures, consider a design that collapses across these values and conduct power analyses using the collapsed variables to address whether that approach improves power. Finally, the table highlights the importance of accurate estimates of correlations between dependent measures when designing for optimal power. In the absence of accurate estimates for correlations between dependent measures, I recommend a conservative approach where the choice of correlations reflects values that limit power. For example, if we had no information about the size of the expected correlations in the present example and were designing to detect a combination of small and strong effects, setting correlations between .1 and .3 provides the most conservative power analysis.

Example 7.3: Multivariate ANOVA

Taking the example used in the previous section, imagine we chose to address whether differences existed across conditions on several dependent measures (i.e., a cross-sectional design) instead of examining change over time. Specifically, we are interested in whether differences exist between groups across different measures of attitudes rather than whether there are differential changes between groups in attitudes over time. This design includes the implicit attitude measure as before but adds paper and pencil measures addressing other aspects of bias (e.g., stereotype endorsement, anxiety, and dislike). MANOVA addresses whether the combination of the dependent measures differs between the two conditions.

Although it may be difficult to estimate correlations between measures, especially for research addressing measures that have not been used together previously, reference to other sources helps establish reasonable estimates. For example, several studies of attitudes toward African Americans used similar dependent measures, so in the absence of information specific to our study targets, relationships found in this work provide some useful estimates of correlations. In examining other studies, correlations between the anxiety, stereotyping, and dislike measures ranged from .35 to .45 (e.g., Tropp & Pettigrew, 2005) and comparatively small correlations (.10) existed between implicit attitudes and the other measures (Aberson & Gaffney, 2009). Table 7.5 shows the pattern of correlations used for the present analysis.

Previous examples discussed the mean differences shown in Table 7.5 for implicit attitudes but did not address the mean differences for the other measures. Earlier, we established a meaningfully sized effect for implicitly held attitudes (see Chapter 6). With regard to the effects for the other variables, we can consider both the size of the implicit effect and content of the experimental

TABLE 7.5 Descriptive Statistics for MANOVA Example

	Implicit	*Stereotype*	*Anxiety*	*Dislike*
Implicit	$\mu_t = 0.0$, $\sigma_t = .40$ $\mu_c = -.25$, $\sigma_c = .40$	$\rho = .10$	$\rho = .10$	$\rho = .10$
Stereotype	$\rho = .10$	$\mu_t = 1.0$, $\sigma_t = 5.0$ $\mu_c = -2.0$, $\sigma_c = 5.0$	$\rho = .35$	$\rho = .45$
Anxiety	$\rho = .10$	$\rho = .35$	$\mu_t = 2.4$, $\sigma_t = 1.6$ $\mu_c = 2.0$, $\sigma_c = 1.6$	$\rho = .40$
Dislike	$\rho = .10$	$\rho = .45$	$\rho = .40$	$\mu_t = -0.7$, $\sigma_t = 1.2$ $\mu_c = -1.0$, $\sigma_c = 1.2$

manipulation. First, since the experimental manipulation focuses on negating stereotypes, it is reasonable to expect a strong effect for stereotype endorsement. For this reason, I set the differences for the stereotyping measures at the same level as the implicit attitude measure. Note that this reflects a moderately strong effect size ($d \approx 0.60$). The liking and anxiety measures are less closely related to the manipulation, so an expectation of smaller effects is reasonable ($d = 0.25$).

As seen in Table 7.4, there are small effects for some variables across condition (anxiety and liking) and moderate to strong effects for others (implicit and stereotyping). The correlations between the dependent measures were set between .10 and .45. The column in Table 7.4 labeled "Small–Moderate" indicates that even if correlations were substantially larger, power would remain relatively constant.

The format of the function is as follows:

```
MANOVA1f(m1.1, m2.1, m3.1, m4.1, m1.2, m2.2, m3.2, m4.2,

s1.1, s2.1, s3.1, s4.1, s1.2, s2.2, s3.2, s4.2,

r1.2_1, r1.3_1, r1.4_1, r2.3_1, r2.4_1, r3.4_1,

r1.2_2, r1.3_2, r1.4_2, r2.3_2, r2.4_2, r3.4_2,

r, s, n, alpha)
```

The values m1.1-m4.1 and m1.2-m4.2 reflect means across the dvs for the first (.1) and second (.2) levels of the between subjects factor. The values s1.1-s4.1 and s1.2-s4.2 reflect standard deviation of the dvs for the first and second levels of the between subjects factor. The *r* values correspond to correlations between the dvs. The _1 or _2 defines the group. The code allows for two to four dvs. The value *n* is overall sample size. Alpha defaults to .05 if no value is entered. The values *r* and *s* should be left blank if providing individuals correlations and standard deviations. For the same standard deviation or correlation across all values, omit the individual values (e.g., s1.1, r1.2_1) and enter a single value for *r* or s.

TABLE 7.6 R Code and Output for Multivariate ANOVA

```
MANOVA1f(n=40, m1.1=0, m2.1=1, m3.1=2.4, m4.1=-0.7, m1.2=-0.25,
m2.2=-2, m3.2=2, m4.2=-1, s1.1=.4, s2.1=5, s3.1=1.6, s4.1=1.2,
s1.2=.4, s2.2=5, s3.2=1.6, s4.2=1.2, r1.2_1=.1, r1.3_1=.1,
r1.4_1=.1, r2.3_1=.35, r2.4_1=.45, r3.4_1=.40, r1.2_2=.1,
r1.3_2=.1, r1.4_2=.1, r2.3_2=.35, r2.4_2=.45, r3.4_2=.40,
alpha=.05)
## [1] "Power MANOVA for n = 40 = 0.8165"
```

TABLE 7.7 R Code and Output for Multivariate ANOVA Examining Different Correlations

```
MANOVA1f(n=40, m1.1=0, m2.1=1, m3.1=2.4, m4.1=-0.7, m1.2=-0.25,
m2.2=-2, m3.2=2, m4.2=-1, s1.1=.4, s2.1=5, s3.1=1.6, s4.1=1.2,
s1.2=.4, s2.2=5, s3.2=1.6, s4.2=1.2, r1.2_1=.1, r1.3_1=.1,
r1.4_1=.1, r2.3_1=-.3, r2.4_1=-.2, r3.4_1=-.2, r1.2_2=.1,
r1.3_2=.1, r1.4_2=.1, r2.3_2=-.3, r2.4_2=-.2, r3.4_2=-.2,
alpha=.05)
## [1] "Power MANOVA for n = 40 = 0.9483"
```

The code and output in Table 7.6 show that 40 participants per group provide adequate power for the MANOVA (.82). If research hypotheses specified rejection of both the multivariate hypothesis and hypotheses for each univariate test, then we would need to investigate power for each univariate ANOVA as well. However, keep in mind that power for detecting multiple effects in the same study is generally lower than power for detecting a single effect (see the discussion of power for multiple tests found in Chapters 5 and 6).

Direction of Correlations in MANOVA

To demonstrate further the influence of patterns of correlations in MANOVA, Table 7.7 shows the analysis from Table 7.6 for a situation where some of the dv expressed negative correlations with each other (stereotyping, anxiety, and liking in this case). When small positive correlations are replaced with small negative correlations while retaining the same pattern of effect sizes, power for the MANOVA jumps from .82 (Table 7.6) to .95 (Table 7.7).

Additional Issues

As mentioned, a common question asked by students and researchers regarding MANOVA is whether reverse coding variables in MANOVA to produce negative correlations improves power. Typically, the question goes something like this: "If I have two positively correlated dv, can I reverse the scale on one of them to yield negatively correlated dv that produce more power?" The short answer is no, this will not affect power.

To demonstrate this, Table 7.8 shows a MANOVA with two variables with a positive correlation (.40). Table 7.9 shows the same variables with the scale for the second variable reversed so that the correlation between the two variables is –.40 and the signs on the means reversed for the second variable in both groups.

Output in Tables 7.8 and 7.9 find the same power estimate (.49). This is because, for the second example, although there is a negative correlation between the predictors, the effects now run in opposite directions. When effects run in opposite directions, the power estimates found in Table 7.9 reverse as shown in Table 7.10.

TABLE 7.8 R Code and Output for Multivariate ANOVA without Reverse Coded Variables

```
MANOVA1f(n=20, m1.1=0, m2.1=1, m1.2=-0.25, m2.2=-2, s1.1=.4,
s2.1=5, s1.2=.4, s2.2=5, r1.2_1=.4, r1.2_2=.4, alpha=.05)
## [1] "Power MANOVA for n=20=0.4879"
```

TABLE 7.9 R Code and Output for Multivariate ANOVA with Reverse Coded Variables

```
MANOVA1f(n=20, m1.1=0, m2.1=-1, m1.2=-0.25, m2.2=2, s1.1=.4,
s2.1=5, s1.2=.4, s2.2=5, r1.2_1=-.4, r1.2_2=-.4, alpha=.05)
## [1] "Power MANOVA for n=20=0.4879"
```

TABLE 7.10 Power as a Function of Effect Size and Correlation Patterns for Effects in Opposite Directions

Corr. Between Measures	*Small–Small* $d = +0.2, -0.2$	*Moderate–Moderate* $d = 0.5, -0.5$	*Strong–Strong* $d = 0.8, -0.8$
.9	.98	1.00	1.00
.8	.80	.99	1.00
.7	.61	.94	1.00
.6	.49	.86	1.00
.5	.40	.77	.99
.4	.34	.69	.98
.3	.30	.61	.96
.2	.26	.55	.93
.1	.24	.50	.90
.0	.22	.46	.87
–.1	.20	.42	.83
–.2	.19	.39	.80
–.3	.18	.36	.76
–.4	.17	.34	.73
–.5	.16	.32	.70
–.6	.15	.30	.67
–.7	.15	.29	.64
–.8	.14	.27	.61
–.9	.14	.26	.59

A final issue when using MANOVA is a clear focus on analysis plans and predictions. Many studies begin with a MANOVA then proceed to univariate ANOVA as follow-up tests on each dv. If a study includes specific predictions regarding individual dependent measures, then I do not see any point in beginning with MANOVA. As stressed in previous chapters, power analyses should address the tests relevant to specific hypotheses.

Summary

Power for ANOVA or LMM designs with between and within subjects factors and MANOVA require estimates of patterns of means, standard deviations for each dependent measure, and the correlation between dependent measures. For both designs, accurate estimates of standard deviation and correlations are particularly important. In addition, careful consideration of the specific test reflecting hypotheses (e.g., omnibus tests vs. tests of trends) is necessary as different types of tests often produce different power estimates. Power for MANOVA is particularly sensitive to estimates of correlations. In the absence of correlation information, the chapter provides guidance for choosing conservative correlation estimates.

8

MULTIPLE REGRESSION

Multiple regression focuses on the prediction of a criterion variable (also known as dependent variable (dv), outcome variable, or response variable) from two or more predictors (also known as independent variables or regressors). The criterion must be continuously scaled. Predictors may be continuously scaled or dichotomous. Predictors with three or more categories are converted to a set of dichotomous predictors via dummy coding (see Cohen, Cohen, West, & Aiken, 2003). This chapter presents power analyses for R^2 *Model*, R^2 *Change*, and regression coefficients in designs using multiple predictors. In addition, the chapter includes tests that examine differences between independent and dependent predictors as well as tests comparing R^2 across independent samples. The chapter also addresses power for detecting multiple effects, how this form of power differs from power to detect individual effects, and the importance of considering the distinction in sample size planning. The Additional Issues section discusses the influence of reliability on power.

Necessary Information

Power analyses for multiple regression focus on the size of meaningful correlations between predictors and the criterion measure. Unlike experimental designs with random assignment where predictors (i.e., factors) are unrelated, predictors in regression analysis often correlate. This correlation between predictors, discussed below as multicollinearity, requires accurate estimation of correlations between predictors to establish realistic estimates of power.

Factors Affecting Power

Several forms of power are of interest for multiple regression. The most common issues are power for the set of all predictors (R^2 *Model*), power for tests of one set

of predictors over another set (R^2 *Change*), and power for a single predictor within a model (regression coefficients). Researchers may seek to address power on some or all forms, depending on their research goals. Additional questions addressed in this chapter involve power for detecting whether predictors are different in size and whether predictors from one sample are stronger than predictors in another sample are. Power considerations differ for each approach but for all tests, larger sample sizes and more liberal alpha increase power.

Many research questions involving regression analysis focus on R^2 *Model*, R^2 *Change*, and tests of coefficients. Power for a set of predictors is tested through estimation of the variance explained by all predictors, termed here R^2 *Model*. The power for the R^2 *Model* is influenced by the amount of variance explained (larger effect size = more power) and the number of predictors. More predictors can lower power because predictors add degrees of freedom to the numerator of the *F* statistic used to test null hypotheses. However, more predictors may increase R^2 *Model* and thus increase power. A small number of predictors that explain a considerable amount of variance are more powerful than a large number of predictors that explain the same amount of variance.

Power for R^2 *Change* involves explanation afforded by addition of a set of predictors over predictors already entered into the prediction model (also known as control variables). This value is of interest for hierarchical multiple regression where the goal is often to address whether addition of a set of variables explains variance over and above existing predictors. Power for change statistics is stronger for predictors that correlate more strongly with the criterion.

Another form of power is for a single predictor within a model. The statistic reflecting this effect is the regression coefficient, either the unstandardized (*b*) or standardized (*b**). This is often called the slope or beta (not to be confused with Type II error). The regression coefficient reflects the strength of the unique relationship between predictor and criterion. That is, what the variable predicts that others cannot. The presentation in this chapter focuses on power for the coefficient. However, the test for the coefficient is equivalent to a test of partial and semipartial correlations for the predictor as well. As with the other approaches, the strength of correlations with the criterion variable affects power. For sets with a single predictor, the power for R^2 *Change* is equivalent to power for the coefficient if the R^2 *Change* reflects the final step of the regression model.

Another form of power is power to detect effects for all of the predictors in the model. As discussed in Chapter 5, Power(All) corresponds to the likelihood of detecting effects across all variables. For example, in the multiple regression context with a three predictor model, Power(All) reflects our ability to find statistically significant results for all three regression coefficients. Power(All) may be substantially smaller than power for individual effects. This form of power is a function of individual power and the correlation between predictors.

Regardless of the test of interest, multicollinearity is a concern. Multicollinearity refers to how strongly the predictor(s) of interest correlate with each other. The

unique explanation afforded by predictors determines the value of R^2 *Change* and the coefficient. The more strongly correlated the predictors, the less unique variance explanation exists, so multicollinearity reduces power. A predictor may be highly correlated with the criterion however, if that predictor correlates strongly with other predictors in the model, the variance it explains over and above the other predictors is limited. In most cases, multicollineary substantially decreases Power(All) as well. For this reason, deriving accurate estimates of the correlations between predictor variables is essential to establishing accurate power estimates.

Other issues covered in this chapter focus on differences between predictors or models. The primary factors affecting power in these cases are the magnitude of the differences between the predictors (or sets of predictors).

Key Statistics

Calculations for R^2 and coefficients are useful for understanding power. Most formulae included here present values for demonstration purposes. The formulae are not necessary for most power calculations but do facilitate an understanding of how multicollinearity influences power. I present formulae for models with two predictors. Adding predictors to a model expands most formulae to increasing levels of complexity such that most texts present only the two predictor formulae.

Formulae for R^2 *and Coefficient Tests*

Formula 8.1 presents the calculation of R^2 for a model with two predictors. The unstandardized coefficient seen in Formula 8.2 technically reflects a population coefficient. Often the symbol beta (β) denotes the unstandardized coefficient for the population. However, β is also used to note the standardized regression coefficient and Type II error. To avoid confusion, I use b to reference the unstandardized population regression coefficient (e.g., Formulae 8.2). I use b^* to represent the standardized coefficient (e.g., Formula 8.3). I use population values for correlations to reflect that in power analysis we make estimates of population values for correlations rather than using values derived from samples.

$$R^2_{y.12} = \frac{\rho^2_{y1} + \rho^2_{y2} - 2\rho_{y1}\rho_{y2}\rho_{12}}{1 - \rho^2_{12}} \tag{8.1}$$

$$b_{y1.2} = \frac{\rho_{y1} - \rho_{y2}\rho_{12}}{1 - \rho^2_{12}} \times \frac{\sigma_y}{\sigma_1}\theta \tag{8.2}$$

$$b^*_{y1.2} = \frac{\rho_{y1} - \rho_{y2}\rho_{12}}{1 - \rho^2_{12}} \tag{8.3}$$

For Formulae 8.1–8.3, *y* refers to the criterion, 1 refers to the first predictor and 2 to the second. For example, *y*.12 means *y* predicted from both 1 and 2; *y*1.2 means *y* predicted from 1 while controlling for predictor 2; ρ_{y2} refers to the correlation between the criterion variable and the second predictor; ρ_{12} reflects the correlation between predictors; and $b_{y1.2}$ is the unstandardized coefficient for the predictor of the criterion by the first predictor while controlling for the second predictor. More simply, this is the coefficient obtained when both predictors are in the same model.

The numerators for each term are a product of the strength of the relationship of interest minus a value that multiplies the correlation between the other predictor and dependent measure by the correlation between predictors. The numerator gets smaller when the predictors overlap more strongly (e.g., predictors positively correlated and predictor–dv relationships in same direction), making for smaller effect sizes. As shown in Formulae 8.1–8.3, under these conditions, the correlation between the predictors (called collinearity with two predictors and multicollinearity with three or more) reduces the size of the R^2 and coefficients. In this way, overlapping predictors limit power.

For designs with a single predictor, Formula 8.4 present the calculation of the unstandardized coefficient (*b*). For these designs, the standardized coefficient (*b**) is equal to the correlation.

$$b_{y1} = \rho_{y1} \times \frac{\sigma_y}{\sigma_1} \tag{8.4}$$

The noncentrality parameter (NCP) for tests of R^2 is the same as for Analysis of Variance (ANOVA) since significance tests use the *F* distribution. Both f^2 and λ may be derived for either the model or the change value, however the approach differs slightly for both tests with regard to the calculation of the effect size (f^2). This effect size estimate is sometimes called partial f^2 as it removes the influence of the other predictors from the denominator (see Formulae 8.5–8.8).

$$f^2_{Model} = \frac{R^2_{Model}}{1 - R^2_{Model}} \tag{8.5}$$

$$\lambda_{Model} = f^2_{Model} df_{error} \tag{8.6}$$

$$f^2_{Change} = \frac{R^2_{Change}}{1 - R^2_{Model}} \tag{8.7}$$

$$\lambda_{Change} = f^2_{Change} df_{error} \tag{8.8}$$

For tests of coefficients, Formulae 8.9–8.12 present calculations of standard errors and the NCP for unstandardized and standardized values. These formulae use variable 1 as an example but maybe adapted to variable 2 by changing references to variable 1 to variable 2 (i.e., change subscripts from 1 to 2). These tests use δ for tests as the NCP. For these tests $\delta^2 = \lambda$.

$$se_{b_1} = \sqrt{\frac{1-R^2_{y.12}}{(1-R^2_{12})*(n-k-1)}} \times \frac{\sigma_y}{\sigma_1} \tag{8.9}$$

$$\delta_{b_1} = \frac{b_1}{se_{b_1}} \tag{8.10}$$

$$se_{b_1^*} = \sqrt{\frac{1-R^2_{y.12}}{(1-R^2_{12})*(n-k-1)}} \tag{8.11}$$

$$\delta_{b_1^*} = \frac{b_1^*}{se_{b_1^*}} \tag{8.12}$$

Formulae for Detecting Differences between Two Independent Coefficients

Tests that address power for detecting differences between two independent coefficients compare coefficients from samples comprised of different people. Analyses comparing a single predictor across two independent samples use Formulae 8.13–8.18. Several formulae use the value "i" to refer to the predictor. We complete this calculation for both coefficients separately with "yi" referring to the relationship between the dependent measures and the predictor of interest. Calculations address either unstandardized (b) or standardized coefficients (*b**).

Calculations first address standard error for each predictor (Formulae 8.13 or 8.16). After calculating the standard error for each of the two *b*s or *b**s (depending on which is used for the analyses), calculate the standard error of the differences (Formulae 8.14 or 8.17). Next, calculate the NCP using Formulae 8.15 or 8.18 (depending on whether using standardized or unstandardized coefficients).

For tests using unstandardized coefficients, refer to Formulae 8.13–8.15.

$$se_{b_i} = \frac{\sigma_y}{\sigma_i}\sqrt{\frac{1}{1-R^2_i}}\sqrt{\frac{1-R^2_y}{n-k-1}} \tag{8.13}$$

$$se_{b_1-b_2} = \sqrt{se^2_{b_1} + se^2_{b_2}} \tag{8.14}$$

$$\delta = \frac{|b_1|-|b_2|}{se_{b_1-b_2}} \tag{8.15}$$

Tests using standardized coefficients refer to Formulae 8.16–8.18.

$$se_{b_1^*} = \sqrt{\frac{1}{1-R^2_i}}\sqrt{\frac{1-R^2_y}{n-k-1}} \tag{8.16}$$

$$se_{b_1^*-b_2^*} = \sqrt{se^2_{b_1^*} + se^2_{b_2^*}} \tag{8.17}$$

$$\delta = \frac{\left|b_1^*\right| - \left|b_2^*\right|}{se_{b_1^* - b_2^*}} \tag{8.18}$$

The calculations represented in Formulae 8.15 and 8.18 test differences in magnitude. That is, whether one predictor is stronger than another is. If you are interested in tests that involve magnitude and direction, simply remove the absolute value symbols.

Formulae for Detecting Differences between Two Dependent Coefficients

Dependent coefficients are those that come from the same analysis. The primary question addressed when comparing dependent coefficients is whether two predictors in the same model differ significantly. Formula 8.19 defines the NCP for this test. The formula requires values from the calculation of the inverse of the correlation matrix (ρ^{ii}, ρ^{jj}, and ρ^{ij}). Cohen et al. (2003) includes calculation details for interested readers, later in the chapter, I provide R code for inverting the matrix and deriving these values.

$$\delta = \frac{\left|b_1\right| - \left|b_2\right|}{\sqrt{se_{b_1}^2 + se_{b_2}^2 - 2se_{b_1}se_{b_2}\left(\frac{\rho^{ij}}{\rho^{ii}\rho^{jj}}\right)}} \tag{8.19}$$

Formulae for Comparing Two Independent R^2 *Values*

A question similar to that addressed by comparing coefficients from independent samples compares R^2 values from different samples. In this case, hypotheses address whether a set of variables predicts more strongly in one analysis than another. Formula 8.20 is appropriate for model or change values.

$$\delta = \frac{\left|R_1^2 - R_2^2\right|}{\sqrt{\left(\frac{4R_1^2\left(1-R_1^2\right)^2\left(n_1-k-1\right)^2}{\left(n_1^2-1\right)\left(n_1+3\right)}\right) + \left(\frac{4R_2^2\left(1-R_2^2\right)^2\left(n_2-k-1\right)^2}{\left(n_2^2-1\right)\left(n_2+3\right)}\right)}} \tag{8.20}$$

Example 8.1: Power for a Two Predictor Model (R^2 Model and Coefficients)

This example focuses on predicting behavioral intentions relevant to affirmative action policies (Intent; y) from two predictors, internal motivation to control prejudice (1) and external motivation to control prejudice (2). Table 8.1 presents correlations between two predictors (internal and external) and a criterion variable (intention). Correlations between the predictors of interest

and the criterion should reflect meaningful observed relationships. That is, what sort of effect would be important to detect. The correlations between predictors should be estimated as accurately as possible, as should predictor–criterion relationships when the researcher is not interested in power for that particular predictor.

Of the three forms of power, my interest here is detecting significant coefficients for internal and external, as well as a significant R^2 *Model*. The variable modern racism, shown in the table is not used for the two predictor example. Another example in the chapter with three predictors makes use of that variable.

Based on commonly observed effect sizes in the affirmative action literature, I determined that a meaningful correlation between each predictor and criterion would have a minimum value of $\rho = .40$. Many predictors of affirmative action beliefs exist so additional variables would have to show moderately large effects to influence on the literature. In short, for the present study, I was not interested in trying to find weak predictors so I set the correlations relatively high. Here, a meaningful relationship is a relatively large one.

Correlations between predictors are not values where it is important to establish the size of a meaningful relationship. For these values, it is more important to have an accurate estimate of the strength of the relationship. That is, how strongly can we expect the predictors to be associated? This is because accurately estimating power for predictor–criterion relationships is dependent on the size of the predictor–predictor correlations. A good source for information when using existing measures are empirical studies that present these correlations, particularly when these relationships are unrelated to focal hypotheses. A scale development study presenting correlations between internal and external motivations (Plant & Devine, 1998) suggested a correlation of –.15 between the two variables.

The distinction between predictor–criterion and predictor–predictor relationships is an important one. When dealing with the criterion variable, focus

TABLE 8.1 Correlations and SDs for Two and Three Predictor Examples

	Intent	*Internal*	*External*	*Modern Racism*
Intent (γ)	$\mu = 1.0$ $\sigma = 7.0$	–	–	–
Internal (1)	$\rho = .40$	$\mu = 1.0$ $\sigma = 1.0$	–	–
External (2)	$\rho = .40$	$\rho = -.15$	$\mu = 1.0$ $\sigma = 1.0$	–
Modern Racism (3)	$\rho = -.40$	$\rho = -.60$	$\rho = .25$	$\mu = 1.0$ $\sigma = 2.0$

on the size of a meaningful relationship. When dealing with correlations between predictors, use the literature (or pretesting) to establish a reasonable estimate. One exception is a design where predictors serve as control variables. For example, when entering a set of variables in the first step of a hierarchical regression analysis and then assessing the influence of one or more variables over and above that set. In that case, the control variables should reflect the accurate instead of meaningful approach (see the three predictor section for an example of this approach).

Calculations based on Formula 8.1, yield R^2 *Model* = .376. One item of interest is that sum of the squared correlations ($.40^2 + .40^2 = .32$) is smaller than R^2 *Model*. This may seem counterintuitive, but this is a product of the direction of the correlations between the predictors. As shown in the following calculation, negative correlations (ρ_{12}) between predictors increase effect sizes, provided that the predictor–dv correlations (ρ_{y1} and ρ_{y2}) are in the same direction (i.e., both negative or both positive). Following the R^2 calculation is calculation of the effect size (Formula 8.5) and NCP (Formula 8.6).

$$R^2_{y.12} = \frac{\rho^2_{y1} + \rho^2_{y2} - 2\rho_{y1}\rho_{y2}\rho_{12}}{1 - \rho^2_{12}}$$

$$= \frac{0.40^2 + 0.40^2 - 2(0.40 * 0.40 * -0.15)}{1 - (-0.15)^2} = 0.376$$

$$f^2_{Model} = \frac{R^2_{Model}}{1 - R^2_{Model}} = \frac{0.376}{1 - .0376} = 0.603$$

$$\lambda_{Model} = f^2_{Model} df_{error} = 0.603(27) = 16.3$$

As demonstrated in Chapter 5, computer approaches allow for calculation of power given λ, *df*, and an $F_{critical}$ value. For example, with a sample of 30 participants, for a test with $\alpha = .05$, with $df_{num} = 2$ (the number of predictors) and $df_{denom} = 27$ (n − Number of predictors − 1), $F_{critical} = 3.35$, and $\lambda = 16.3$, power is .94.

The line of code below calculates power:

```
1-pf(3.35, 2, 27, 16.3)
```

Table 8.2 presents R code and output for conducting power analyses for R^2 *Model* and the coefficient. As in other chapters, the R functions require only the descriptive statistics. The primary information for entry are the correlations. Estimates of the *M* and *SD* can be set at arbitrary values for most analyses. The standard deviation affects the coefficient, but not the power analysis for the coefficient. This is because the ratio of the coefficient to its standard error is a function of the correlations.

The format of the function is as follow:

```
MRC(ry1, ry2, r12, n, alpha, my, m1, m2, sy, s1, s2)
```

Values noted as *y* indicate the dv. Those noted with numbers are predictors. ry1 and ry2 are the correlations between the predictor and dv. r12 is the correlation between predictors. *n* is sample size. Alpha defaults to .05 if not entered. The means and standard deviation are represented with *m* and *s*. These values default to means of 0 and standard deviations of 1. The function handles up to five predictors.

Table 8.2 includes the output relevant to power for R^2 *Model*. With a sample of 30 participants, given the correlations presented in Table 8.1, power is 94%. Keep in mind that power for the R^2 *Model* does not necessarily suggest the same level of power to detect effects for both coefficients in the model.

Tests of coefficients involve calculation of the coefficient, its standard error, and the NCP. As an example, I present calculation of one the unstandardized coefficients, standard error, and the NCP using Formulae 8.2, 8.9, and 8.10.

$$b_{y1.2} = \frac{\rho_{y1} - \rho_{y2}\rho_{12}}{1-\rho_{12}^2} \times \frac{\sigma_y}{\sigma_1} = \frac{0.40 - 0.40(-0.15)}{1-(-0.15)^2} \times \frac{7}{1} = 3.294$$

$$se_{b_1} = \sqrt{\frac{1-R_{y.12}^2}{(1-R_{12}^2)*(n-k-1)}} \times \frac{\sigma_y}{\sigma_{x_1}}$$

$$= \sqrt{\frac{1-0.376}{(1-0.15^2)*(30-2-1)}} \times \frac{7}{1} = 1.076$$

$$\delta_{b_1} = \frac{b_1}{se_{b_1}} = \frac{3.294}{1.076} = 3.061$$

A sample of 30 participants, for a test with $\alpha = .05$, $df = 27$ (n – # predictors – 1) yields $t_{critical} = 2.05$. With $\delta = 3.061$ (alternatively, we can square this value to produce $\lambda = 9.37$), the line of code below calculates power as .84.

```
1-pt(2.05, 27, 3.061)
```

Table 8.2 also shows power analysis for the coefficients for each predictor. This comes from the same analysis produced by the code in Table 8.2. With a sample of 30, power is around 84% for both predictors. These values are equal as both predictor–dv correlations were .40. Although power is good for both predictors, power for the coefficients are less than the power for R^2 *Model*.

TABLE 8.2 R Code and Output for Two Predictors

```
MRC(ry1=.40, ry2=.40, r12=-.15, n=30)
## [1] "Sample size is 30"
## [1] "Power R2 = 0.937"
## [1] "Power b1 = 0.839"
## [1] "Power b2 = 0.839"
```

Example 8.2: Power for Three Predictor Models

The three predictor model expands on Example 8.1 through addition of a third predictor (modern racism). In this example, one additional test is power for R^2 *Change* for a set including internal and external after considering the influence of modern. Set analyses of this sort would be the preferred approach for dummy-coded predictors (e.g., a predictor with three categories coded into two dichotomous variables then entered as a set). Also addressed is power for coefficients within a three predictor model.

This analysis controls for the influence of modern racism, so this is not a test where we are interested in a meaningful relationship between this predictor and the criterion per se. This analysis investigates the influence of internal and external over modern because modern racism is an established predictor of affirmative action relevant beliefs that may be correlated with internal and external motivations to control prejudice. Estimates for the modern racism variable correlations come from two sources. Information from meta-analyses (Harrison et al., 2006) and a scale development article (Plant & Devine, 1998), suggested correlations for modern racism and the other variables found in Table 8.1.

The format of the R2ch function is:

R2ch(ry1, ry2, ry3, r12, r13, r23, n, alpha)

The inputs for this function are the same as the MRC function.

Table 8.3 shows that we need 24 participants for power≥.80 for R^2 *Change*. In this example, small increases in sample size influence power considerably. This result focuses on R^2 *Change* for a set of variables. Researchers often desire an outcome wherein not only was R^2 *Change* significant but also individual contribution of each variables within that set were significant. Addressing this issue requires power analysis for the coefficients for our predictors (the internal and external motivation variables). Keep in mind that the sample size yielding adequate power for the model and the set (change) will not necessarily yield high power for each coefficient.

Table 8.4 shows that with $n = 24$, the power for the coefficients (.853 for external and .215 for internal) are divergent. The output also shows power for R^2 *Model* (.90). Power for the coefficients diverge because of differences in the predictor's correlation with modern racism (the control variable). Internal

TABLE 8.3 R Code and Output for R^2 *Change* Power ($n = 24$)

```
R2ch(ry1=.40, ry2=.40, ry3=-.40, r12=-.15, r13=-.60, r23=.25,
n=24)
## [1] "R2 Model = 0.4667"
## [1] "R2 Change Vars2 and 3 over Var1 = 0.3067, Power = 0.8097"
## [1] "R2 Change Vars1 and 3 over Var2 = 0.3067, Power = 0.8097"
## [1] "R2 Change Vars1 and 2 over Var3 = 0.3067, Power = 0.8097"
```

TABLE 8.4 R Code and Output for Three Predictors ($n = 24$)

```
MRC(ry1=.40, ry2=.40, ry3=-.40, r12=-.15, r13=-.60, r23=.25,
n=24)
## [1] "Sample size is 24"
## [1] "Power R2 = 0.904"
## [1] "Power b1 = 0.215"
## [1] "Power b2 = 0.853"
## [1] "Power b3 = 0.417"
```

TABLE 8.5 R Code and Output for Coefficient Power ($n = 110$)

```
MRC(ry1=.40, ry2=.40, ry3=-.40, r12=-.15, r13=-.60, r23=.25,
n=110)
## [1] "Sample size is 110"
## [1] "Power R2 = 1"
## [1] "Power b1 = 0.798"
## [1] "Power b2 = 1"
## [1] "Power b3 = 0.987"
```

motivation had a stronger relationship with modern racism, as a result internal motivation is a weaker unique predictor of intentions.

To achieve power of .80 (or more) for both internal and external coefficients requires a larger sample. Table 8.5 presents an analysis producing the desired level of power for the coefficients.

The differences between sample size requirements for tests of the model, change, and coefficients highlight important considerations in multiple regression designs. First, different research questions correspond to different power estimates. Researchers should first decide on the question that is most relevant then design for appropriate power for that question. Second, multicollinearity influences power. When predictors are highly correlated, power for coefficients drop considerably. Tests of coefficients examine variability explained uniquely by a predictor. The more strongly predictors correlate, the less unique variance there is to go around. This can be seen by examining power in the three predictor design. External motivation had small correlations with the other predictors but internal had a strong correlation with modern. External has greater power because it explains more variance that the other variables cannot account for.

Example 8.3: Power for Detecting Differences between Two Dependent Coefficients

This section deals with determining if one predictor in a model is significantly stronger than another predictor in the same model. For example, an analysis designed to test whether internal motivation was a stronger predictor than external motivation within Example 8.2, involves dependent coefficients as the

coefficients come from the same participants. It is tempting to think that if one predictor is statistically significant and the other is not that this would mean that one predictor is stronger than the other is. However, imagine an analysis where one coefficient in the model was barely significant (e.g., $p = .049$) and the other missed the mark (e.g., $p = .051$). In this case one predictor is not likely stronger than the other.

For Example 8.2, imagine that we wanted to design to conclude that internal motivations were stronger predictors of intentions than external motivations. Using the data from Table 8.1, output from Table 8.6 to obtain the necessary beta and se values, and some additional calculations allows for determination of the NCP using Formula 8.19.

To get values for calculation, I used the MRC_shortcuts function.[1] The format of the MRC_shortcuts function is:

MRC_shortcuts(ry1, ry2, ry3, r12, r13, r23, n, my, m1, m2, m3, sy, s1, s2, s3)

The format of the function is the same as for MRC.

One aspect of this calculation that deserves special mention are the values noted ρ^{ii}, ρ^{ij}, and ρ^{ij}. These values come from the inverse of the correlation matrix. For those with matrix algebra backgrounds, the calculation is simple. For those without, computer protocols easily accomplish these calculations.

Table 8.6 provides values for the coefficients, yielding $b_1 = 3.73$ and $b_2 = 1.75$ with the corresponding standard errors of $SE_{b1} = 0.513$ and $SE_{b2} = 0.621$. Based on the inverted matrix of predictor correlations (calculated by the computer[2]) $\rho^{ii} = 1.56$, $\rho^{ij} = 0.00$, and $\rho^{ij} = 1.07$. Using Formula 8.19, produces $\delta = 2.46$.

$$\delta = \frac{|b_1| - |b_2|}{\sqrt{se_{b_1}^2 + se_{b_2}^2 - 2se_{b_1}se_{b_2}\left(\frac{\rho^{ij}}{\rho^{ii}\rho^{jj}}\right)}}$$

$$= \frac{3.73 - 1.75}{\sqrt{0.513^2 + 0.621^2 - 2*0.513*0.621\left(\frac{0}{1.56*1.07}\right)}} = 2.46$$

For a sample of $n = 110$, plug $\delta = 2.46$ with $df = 106$ (n minus the total number of predictors minus 1; 110–3 –1) and a $t_{critical}$ value of 1.98 (.05, 2-tailed) into the code line below. Another approach is to square delta to get λ. Either approach yields Power = .68.

Using t and δ: 1-pt(1.98, 106, 2.46)

Using F and λ: 1-pf(3.93, 1,106, 6.05)

The R code in Table 8.7 conducts these tests but does not require the calculations. The code makes all possible coefficient comparisons. The depb function

TABLE 8.6 Output for Dependent Coefficients Calculation Example

```
MRC_shortcuts(ry1=.40, ry2=.40, ry3=-.40, r12=-.15, r13=-.60,
r23=.25, n=110, my=1, m1=1, m2=1, m3=1, sy=7, s1=1, s2=1, s3=2)
##
## Call:
## lm(formula = X1~X2+X3+X4, data = pop2)
##
## Residuals:
## Min 1Q Median 3Q Max
## -11.4388 -3.4470 -0.1457 3.4365 15.6208
##
## Coefficients:
##             Estimate   Std. Error   t value   Pr(>|t|)
## (Intercept)   -3.1417       1.0762   -2.919   0.00429**
## X2             1.7500       0.6207    2.820   0.00574**
## X3             3.7333       0.5128    7.280   6.11e-11***
## X4            -1.3417       0.3169   -4.234   4.90e-05***
## -
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.184 on 106 degrees of freedom
## Multiple R-squared: 0.4667, Adjusted R-squared: 0.4516
## F-statistic: 30.92 on 3 and 106 DF, p-value: 1.939e-14
```

takes the same form as MRC but requires only correlations, sample size, and alpha. The code produces power for all coefficient comparisons.

The format of the depb function is:

depb(ry1, ry2, ry3, r12, r13, r23, n, alpha)

Table 8.7 shows power for $n = 110$, is about .68. A sample of 143 would achieve 80% power for the test of differences.

TABLE 8.7 R Code and Output for Comparing Dependent Coefficients

```
depb(ry1=.40, ry2=.40, ry3=-.40, r12=-.15, r13=-.60, r23=.25,
n=110, alpha=.05)
## [1] "Sample size is 110"
## [1] "Power Comparing b1 and b2 = 0.685"
## [1] "Power Comparing b1 and b3 = 0.161"
## [1] "Power Comparing b2 and b3 = 0.243"
depb(ry1=.40, ry2=.40, ry3=-.40, r12=-.15, r13=-.60, r23=.25,
n=143, alpha=.05)
## [1] "Sample size is 143"
## [1] "Power Comparing b1 and b2 = 0.8"
## [1] "Power Comparing b1 and b3 = 0.197"
## [1] "Power Comparing b2 and b3 = 0.304"
```

Example 8.4: Power for Detecting Differences between Two Independent Coefficients

Another possible question is whether a predictor is stronger for one group than another. For example, in the internal and external motivation example, a reasonable argument might be that a sample of college students might respond more strongly to external motivations than an older, non-college sample.

Example 8.4 uses the data from the earlier example to represent a college student sample and adds a second set of values to represent a non-college (adult) sample. Table 8.8 shows these values.

The basic approach is to take the difference between the unstandardized regression coefficients divided by the standard error of the differences between coefficients using Formulae 8.15 or 8.18. As with other aspects of analyses in this chapter, I use R code to derive several values (i.e., skipping hand calculations). The function called mrc_short2 requires the same input as the indb function that is detailed later in the chapter.

For calculations, I use R code (MRC_short2) to derive coefficients, R^2_y, and R^2_i for calculations and/or input into the power analysis code. For each of the two samples, we derive R^2 for prediction of the dv. Tables 8.9 presents output from these analyses. As there is a considerable amount of output, the relevant values are bolded.

The output in Table 8.9 shows $R^2_y = .467$, reflecting how well the criterion variable is predicted by predictors in the student sample. Also from this output, we take $b_1 = 3.733$, the unstandardized coefficient for external. Table 8.9 also presents output for R^2_i. From the table, $R^2_i = .063$, reflecting how well the predictor of interest (external) is predicted by the other predictor variables (again for the student sample). Since the predictors show the same correlations across both samples, I present this analysis only once (it produces $R^2_i = .063$ for both samples). Using these values and Formula 8.13 we can calculate the standard error of each coefficient (shown only for the college sample).

$$se_{b_1} = \frac{\sigma_y}{\sigma_1}\sqrt{\frac{1}{1-R_i^2}}\sqrt{\frac{1-R_y^2}{n-k-1}} = \frac{7}{1}\sqrt{\frac{1}{1-0.063}}\sqrt{\frac{1-0.467}{50-3-1}} = 0.778$$

TABLE 8.8 Correlations for Both Populations with Student Sample on Lower Diagonal and Adult Sample on Upper Diagonal

	Intent	*Internal*	*External*	*Modern*
Intent (y)	$\sigma = 7.0$	$\rho = .40$	$\rho = .10$	$\rho = -.40$
Internal (1)	$\rho = .40$	$\sigma = 1.0$	$\rho = -.15$	$\rho = -.60$
External (2)	$\rho = .40$	$\rho = -.15$	$\sigma = 1.0$	$\rho = .25$
Modern Racism (3)	$\rho = -.40$	$\rho = -.60$	$\rho = .25$	$\sigma = 2.0$

Note that in Table 8.9, the standard error for external is 0.778, just as found in the calculation. The values for the adult sample are $b_2 = 1.493$ and $se_{b2} = 0.928$. These values and those from the student sample, allow for calculation of the standard error of the differences (Formula 8.14) and then the NCP (Formula 8.15).

$$se_{b_1 - b_2} = \sqrt{se_{b_1}^2 + se_{b_2}^2} = \sqrt{0.778^2 + 0.928^2} = 1.211$$

$$\delta = \frac{|b_1| - |b_2|}{SE_{b_1 - b_2}} = \frac{3.733 - 1.493}{1.211} = \frac{2.24}{1.211} = 1.85$$

Power for $\delta = 1.85$, using $t_{critical}$ for a two-tailed test with $\alpha = .05$ and $df = 92$ (total sample size − # of predictors in first model − # of predictors in second model −2) comes to .45.

TABLE 8.9 Output for Independent Coefficients Calculation Example

```
MRC_short2(ry1_1=.40, ry2_1=.40, ry3_1=-.40, r12_1=-.15,
r13_1=-.60, r23_1=.25, ry1_2=.40, ry2_2=.10, ry3_2=-.40,
r12_2=-.15, r13_2=-.60, r23_2=.25, n1=50, n2=50, alpha=.05,
my_1=1, m1_1=1, m2_1=1, m3_1=1, sy_1=7, s1_1=1, s2_1=1,
s3_1=2, my_2=1, m1_2=1, m2_2=1, m3_2=1, sy_2=7, s1_2=1,
s2_2=1, s3_2=2)
## [1] "Overall Analyses for R2 Full model and coefficients,
First Group"
##              Estimate   Std. Error   t value   Pr(>|t|)
## (Intercept)  -3.1417    1.6318       -1.925    0.06040
## X2            1.7500    0.9422        1.857    0.06966
## X3            3.7333    0.7785        4.796    1.74e-05***
## X4           -1.3417    0.4810       -2.789    0.00766**
## -
## Multiple R-squared: 0.4667, Adjusted R-squared: 0.4319
## [1] "Analyses for R2i (how well predictor is explained by
other predictors, First Group)"
## Coefficients:
##              Estimate   Std. Error   t value   Pr(>|t|)
## (Intercept)  8.750e-01  2.779e-01    3.149     0.00285**
## X2           -3.172e-16 1.765e-01    0.000     1.00000
## X4           1.250e-01  8.827e-02    1.416     0.16334
## Multiple R-squared: 0.0625, Adjusted R-squared: 0.02261
## [1] "Overall Analyses for R2 Full model and coefficients,
Second Group"
## Coefficients:
##              Estimate   Std. Error   t value   Pr(>|t|)
## (Intercept)  -1.1817    1.9446       -0.608    0.5464
## X2            1.7500    1.1227        1.559    0.1259
## X3            1.4933    0.9276        1.610    0.1143
## X4           -1.0617    0.5732       -1.852    0.0704
## Multiple R-squared: 0.2427, Adjusted R-squared: 0.1933
```

TABLE 8.10 R Code and Output for Comparing Independent Coefficients

```
indb(ry1_1=.40, ry2_1=.40, ry3_1=-.40, r12_1=-.15, r13_1=-.60,
r23_1=.25, ry1_2=.40, ry2_2=.10, ry3_2 =-.40, r12_2=-.15,
r13_2=-.60, r23_2=.25, n1=50, n2=50, alpha=.05)
## [1] "Sample size Group 1 = 50 Group 2 = 50"
## [1] "Power Comparing b1 across samples = 0.05"
## [1] "Power Comparing b2 across samples = 0.449"
## [1] "Power Comparing b3 across samples = 0.066"
indb(ry1_1=.40, ry2_1=.40, ry3_1=-.40, r12_1=-.15, r13_1=-.60,
r23_1=.25, ry1_2=.40, ry2_2=.10, ry3_2=-.40, r12_2=-.15, r13_2=-
.60, r23_2=.25, n1=115, n2=115, alpha=.05)
## [1] "Sample size Group 1 = 115 Group 2 = 115"
## [1] "Power Comparing b1 across samples = 0.05"
## [1] "Power Comparing b2 across samples = 0.816"
## [1] "Power Comparing b3 across samples = 0.089"
```

Table 8.10 provides R code and output for the analysis used in hand calculations. The format of the function (indb) is as follows:

indb(ry1_1, ry2_1, ry3_1, r12_1, r13_1, r23_1, n1,

ry1_2, ry2_2, ry3_2, r12_2, r13_2, r23_2, n2, alpha = .05)

The values in the function are the same as for MRC with _1 and _2 used to designate the group.

As shown in the earlier calculation, the output in Table 8.10 indicates that $n = 50$ per group produces Power = .45. For $n = 115$ per sample, power is .82.

Example 8.5: Comparing Two Independent R^2 Values

Another option with regression is to compare overall prediction across samples. For example, if our interest was to test if our ability to predict intentions was significantly better in the student sample compared to the adult sample, this sort of question involves comparison of the R^2 *Model* values for each analysis.

This example examines the R^2 for both samples. Although the example presents comparisons of R^2 *Model*, this approach works for R^2 *Change* comparisons as well. Again, the shortcut code used in the previous example does this calculation easily, producing R^2 for the student sample of .467 and .243 for the adult sample taken from Table 8.9. Table 8.9 presents R^2 for each sample. Apply these values to Formula 8.20 to calculate the NCP. The example begins with $n = 115$ for each group (230 overall), reflecting the sample size found for differences between coefficients as a starting point.

TABLE 8.11 R Code and Output for Comparing Two Independent R^2s

```
indR2(ry1_1=.40, ry2_1=.40, ry3_1=-.40, r12_1=-.15, r13_1=-.60,
r23_1=.25, ry1_2=.40, ry2_2=.10, ry3_2=-.40, r12_2=-.15,
r13_2=-.60, r23_2=.25,
n1=115, n2=115, alpha=.05)
## [1] "Power=0.672, n1 = 115, n2 = 115, LLdiff = 0.041,
ULdiff=0.407"
indR2(ry1_1=.40, ry2_1=.40, ry3_1=-.40, r12_1=-.15, r13_1=-.60,
r23_1=.25, ry1_2=.40, ry2_2=.10, ry3_2=-.40, r12_2=-.15,
r13_2=-.60, r23_2=.25, n1=160, n2=160, alpha=.05)
## [1] "Power=0.801 n1=160, n2=160, LLdiff=0.067, ULdiff=0.381"
```

$$\delta = \frac{\left|R_1^2 - R_2^2\right|}{\sqrt{\left(\frac{4R_1^2\left(1-R_1^2\right)^2\left(n_1-k-1\right)^2}{\left(n_1^2-1\right)\left(n_1+3\right)}\right)+\left(\frac{4R_2^2\left(1-R_2^2\right)^2\left(n_2-k-1\right)^2}{\left(n_2^2-1\right)\left(n_2+3\right)}\right)}}$$

$$= \frac{\left|0.467-0.243\right|}{\sqrt{\left(\frac{(4\times 0.467)(1-0.467)^2(115-3-1)^2}{(115^2-1)(115+3)}\right)+\left(\frac{(4\times 0.243)(1-0.243)^2(115-3-1)^2}{(115^2-1)(115+3)}\right)}}$$

$$= 2.43$$

Evaluating power for $\delta = 2.43$, requires a *t*-critical value for two-tailed test with $\alpha = .05$ and $df = 222$ (that is total sample size − # of predictors in first model − # of predictors in second model −2). Using the procedures detailed earlier, $\delta = 2.43$ corresponds to Power = .67.

Table 8.11 provides code and output for comparing independent R^2 values. Power estimates deviate slightly from hand calculations due to rounding. The format of the function is as follows:

indR2(ry1_1, ry2_1, ry3_1, r12_1, r13_1, r23_1, n1,

ry1_2, ry2_2, ry3_2, r12_2, r13_2, r23_2, n2, alpha = .05)

The input for the function is identical to indb shown in the previous section.

Table 8.11 shows that for this analysis, power of .80 requires a sample of 320 overall, given that there are equal numbers of participants in each group.

Multiplicity and Direction of Predictor Correlations

As discussed in Chapter 5, designs involving two or more predictors require different conceptualizations of power. In ANOVA, a researcher might only

have an interest in detecting a single effect (i.e., the interaction). However, in multiple regression designs, researchers commonly want to detect significant coefficients for all of the predictors in the model. Applications of power analyses for designs with multiple predictors typically yield an estimate of power for each predictor, but not power to detect all of them in the same study. Problematically, power to detect multiple effects differs considerably from power to detect individual effects. In most situations, power to detect multiple effects is considerably lower than the power for individual effects. The lack of attention to this form of power is a likely source underpowered research in the behavioral sciences (Maxwell, 2004).

Inflation of Beta Error

Power is reduced in designs that aim to detect significant effects for multiple predictor variables through inflation of the familywise beta error rate (Maxwell, 2004 for a technical discussion). This issue is similar to inflation of α or Type I error. When conducting multiple tests, Type I error rates for the family of tests (a.k.a., familywise alpha) rise as a function of alpha and number of tests conducted. Equation 8.21 provides an estimate of familywise α error for multiple comparisons. With three tests using a pairwise alpha (α_{pw}) of .05, familywise alpha (α_{fw}) is .14.

$$\alpha_{fw} = 1 - (1 - \alpha_{pw})^c \tag{8.21}$$

The same process occurs for the familywise probability of making a β or Type II error (Equation 8.22). I refer to the familywise β as β_{fw}. In a study designed to produce $\beta = .20$ (called β_{ind} for beta individual) for each of its three predictors (i.e., Power = .80 for each predictor), the likelihood of a single β error among those tests is higher than the .20 Beta error rate for the individual tests. The β_{fw} value converts to power to detect all of the effects in the design by taking $1-\beta_{fw}$. I refer to this value as Power(All).

$$\beta_{fw} = 1 - (1 - \beta_{ind})^c \tag{8.22}$$

Table 8.12 shows β_{fw} and Power(All) for two through 10 predictors. The table includes results for design with Power = .80 and Power = .95 for each individual predictor. The difference between individual power suggests that more predictors make it unlikely to find significant for every effect. This table is useful for a conceptual understanding of β_{fw}, however these results (and Formula 8.22) are only accurate for calculations where all tests have the same power. More importantly, correlations between predictors dramatically influences Power(All).

TABLE 8.12 Familywise Type II Error (Beta) Rates for Predictors using $\beta_{pw} = .20$ (Power = .80)

Number of Predictors	*Power = .80*		*Power = .95*	
	β_{fw}	*Power(All)*	β_{fw}	*Power(All)*
2	.360	.640	.098	.903
3	.488	.512	.143	.857
4	.590	.410	.185	.815
5	.672	.328	.226	.774
6	.738	.262	.265	.735
7	.790	.210	.302	.698
8	.832	.168	.337	.663
9	.866	.134	.370	.630
10	.893	.107	.401	.599

Note
All predictors uncorrelated. This table is not accurate for correlated predictors.

Power(All) for Designs with Correlated Predictors

Calculation of β_{pw} and Power(All) is simple when predictors are uncorrelated. However, predictors usually correlate to some degree in multiple regression applications. The influence of correlations between predictors and Power(All) is a function of the strength and direction of correlations between predictors.[3] When predictors correlate positively with each other, Power(All) decreases. When predictors correlate negatively, Power(All) increases.

Table 8.13 shows how predictor correlations influence Power(All) for two predictor models. Power for each predictor is set at .80 (the size of the correlations between the predictors and the dv, listed as "required correlations" change to obtain Power = .80). The Reject All column reflects Power(All) estimates derived by simulation of 10,000 samples drawn from a population with the given correlations. The range of values for Power(All) is roughly .59 to .72 with more power generated as correlations between predictors move from strongly positive to strongly negative. Since this approach involves simulation, there is deviation from the theoretical probabilities. For example, Power(All) for two predictors with Power = .80 and no correlation between predictors is theoretically .64 but is it .6348 in the simulation.

Table 8.13, suggests that negative correlations between predictors are advantageous. However, it is unlikely to find predictors that correlate strongly in the negative direction with each other when both predictors have a consistent (i.e., all positive or all negative) relationship with the dv. Situations consistent with the positive correlation results in the table are far more common.

Table 8.14 provides Power(All) for three predictors. In each situation, Power = .80 for the individual predictors and sample size is 100. Despite

TABLE 8.13 Power(All) for Two Predictors with Power = .80 and Varying Levels of Correlation

Correlation Between Predictors	*Required x–y Correlations*	*Reject None*	*Reject One*	*Reject All*
–.80	.1274	.1294	.1492	.7214
–.60	.1891	.1074	.2029	.6897
–.40	.2445	.0816	.2458	.6726
–.20	.2999	.0564	.2912	.6524
.00	.3594	.0463	.3189	.6348
.20	.4266	.0279	.3518	.6203
.40	.5070	.0190	.3708	.6102
.60	.6102	.0102	.3864	.6034
.80	.7561	.0033	.4107	.5860

Note
Required *x–y* correlation is the correlation between each predictor and the dv to produce Power = .80 with $n = 50$.

substantial power for individual predictors, Power(All) can be as low as .44 for a model with strongly correlated predictors. As with two predictor models, Power(All) increases as predictor correlations move from positive to negative. However, Power(All) is smaller with more predictors.

Some of the values in Table 8.14, represented as n/a, are not possible. For example, there is no predictor–dv correlation where it is possible to have correlations of –.60 or –.80 between the predictors (given $n = 100$).

Given these issues, I offer several recommendations for designs where the goal is to detect multiple effects. First, whenever possible use uncorrelated or

TABLE 8.14 Power(All) for Three Predictors with Power = .80 and Varying Levels of Correlation

Correlation Between Predictors	*Required x–y Correlations*	*Reject None*	*Reject One*	*Reject Two*	*Reject All*
–.80	n/a	–	–	–	–
–.60	n/a	–	–	–	–
–.40	.0804	.0793	.1030	.1800	.6377
–.20	.1692	.0268	.1129	.3046	.5557
.00	.2583	.0091	.1005	.3678	.5226
.20	.3569	.0033	.0892	.4251	.4824
.4	.4703	.0008	.0678	.4681	.4633
.6	.6057	.0001	.0506	.5000	.4493
.8	.7747	.0000	.0435	.5211	.4354

Note
Required *x–y* correlation is the correlation between each predictor and the dv to produce Power = .80 with $n = 100$.

slightly negatively correlated predictors. If predictors demonstrate strong positive correlations, recognize that this increases sample size requirements. Second, consider factor/components analyses to help identify uncorrelated predictors. Third, if your goal is to have Power(All) = .80, design for greater power on individual predictors. Example 8.6 demonstrates tools for such analyses. Finally, consider set analyses for highly correlated predictors. In general, test of R^2 *Change* for the set are more powerful than tests of the unique contribution of each predictor when the predictors are highly correlated.

Example 8.6: Power(All) with Three Predictors

The example in the table demonstrates use of the MRC_all function. The format of the function is:

MRC_all(ry1, ry2, ry3, r12, r13, r23, n, my, m1, m2, m3, sy, s1, s2, s3, rep)

The input for the function is identical to that for MRC with one exception. Power produced by this function involves simulation samples. The value rep defines the number of simulations of size *n* from a large population (100,000 cases). As a default, rep is set at 10,000. For this reason, the code may take a minute or two to run (particularly on slower computers). To run faster (e.g., to get a quick estimate before running a full analysis) reduce the number of reps.

TABLE 8.15 Power(All) for Three Predictors

```
MRC_all(ry1=.50, ry2=.50, ry3=.50, r12=.2, r13=.3, r23=.4, n=82)
## [1] "Sample size is 82"
## [1] "Power R2=1"
## [1] "Power b1=0.9758"
## [1] "Power b2=0.9354"
## [1] "Power b3=0.8012"
## [1] "Proportion Rejecting None=0"
## [1] "Proportion Rejecting One=0.007"
## [1] "Proportion Rejecting Two=0.2736"
## [1] "Power ALL (Proportion Rejecting All)=0.7194"
MRC_all(ry1=.50, ry2=.50, ry3=.50, r12=.2, r13=.3, r23=.4, n=94)
## [1] "Sample size is 94"
## [1] "Power R2=1"
## [1] "Power b1=0.9888"
## [1] "Power b2=0.9639"
## [1] "Power b3=0.852"
## [1] "Proportion Rejecting None=0"
## [1] "Proportion Rejecting One=0.0022"
## [1] "Proportion Rejecting Two=0.1909"
## [1] "Power ALL (Proportion Rejecting All)=0.8069"
```

Table 8.15 represents a situation where the predictors all share .5 correlations with the dv and correlations of .2, .3, and .4 with each other. A sample of 82 yields a minimum power of .80 for each predictor. However, the power to detect significance for all the coefficients in the same design is only .72 for detecting all of the effects in the same design. To obtain Power(All) = .80 requires a sample of 94 participants. This may seem like a small number but it represents a 15% increase in sample size.

Additional Issues

Reliability

Reliability plays a major role in regression analysis with continuously scaled variables. Less reliable measures reduce the size of correlations observed in samples (e.g., Hunter & Schmidt, 1994). Since poor reliability attenuates observed relationships, less reliable measures produce smaller effect sizes and reduce power. For example, two variables might have a .60 correlation in the population (ρ_{true}); however, unreliable measures may reduce the observed correlation (ρ_{obs}). Formula 8.21 show how reliability influences the observed correlation (α_1 is the reliability for the predictor, α_y is for the criterion).

$$\rho_{obs} = \rho_{true}\sqrt{\alpha_1\alpha_y} \tag{8.21}$$

In the example below both the variables demonstrate mediocre reliability ($\alpha_1 = \alpha_y = .50$). In this case, the observed correlation is half the size of the population correlation. Of course, if the effect size observed is considerably smaller than the expected effect in the population (i.e., the value used in power analysis), power falls.

$$0.30 = 0.6\sqrt{0.5 * 0.5}$$

The next calculation shows a situation where both variables have strong reliabilities ($\alpha_1 = \alpha_y = .90$). In this case, the observed correlation is closer to the population value.

$$0.54 = 0.6\sqrt{0.9 * 0.9}$$

Reliability is also important for experimental designs, but often to a lesser extent. For experimental designs factors based on random assignment are considered perfectly reliable (i.e., $\alpha_1 = 1.0$) so the influence of reliability on observed relationships comes only from the dv rather than the dv and the factor.

Summary

This chapter presented power for R^2 *Model*, R^2 *Change*, and coefficients for multiple regression. For these tests the primary information required are the correlations between variables or alternatively, estimates of R^2. Estimates of correlations between predictors and the dependent measure (as well as R^2) should reflect meaningful levels of association whereas estimates for correlations between predictors focus on accuracy. This chapter also presented tests for comparisons between independent coefficients, dependent coefficients, and independent R^2 values. Each test requires estimates of the value of interest and correlations between all predictors and/or R^2 values. Finally, the chapter focused on detecting all effects in a model. In general, power to detect all effects is smaller than power for individual effects.

Notes

1. The MRC_shortcuts function can be used to obtain R^2 and coefficient values that correspond to power analyses.
2. r12<-.15; r13←.60; r23<-.25
   mat<-cbind(c(1, r12, r13),c(r12, 1, r23),c(r13, r23, 1))
   inv<-solve(mat)*mat
   # 1 vs 2
   pij1<-inv[1,2] #inv of cor between pred of interest 1 vs 2
   pii1<-inv[1,1] #inv of cov, v1
   pjj1<-inv[2,2] #inv of cov, v2
3. I discuss these issues in terms of the directions of the correlations between variables. Maxwell (2004) focuses on correlations between *coefficients*. A positive correlation between predictors would reflect a negative relationship among coefficients (i.e., as one coefficient rises the other tends to fall).

9

ANALYSIS OF COVARIANCE, MODERATED REGRESSION, LOGISTIC REGRESSION, AND MEDIATION

This chapter examines ANCOVA, regression designs with interactions, logistic regression (LR), and indirect effects (mediation). Some analyses in the chapter expand work in Chapters 6–8, and in some cases continue examples from those chapters. Additional issues include reliability influences on detection of regression interactions. Few new formulae are presented as calculations of some effect sizes and other values for the analyses are outside the scope of this text. For details on these calculations, see the work of Tabachnick and Fidell (2007a; 2007b), Cohen, Cohen et al. (2003), Aguinis (2004), Menard (2009), and Hayes (2017).

Analysis of Covariance

Necessary Statistics

Covariate analyses require means and standard deviations for each group or cell (as with ANOVA designs) and estimates of correlations between the covariate and dependent measure (as with regression).

Factors Affecting Power

Inclusion of covariates often increases power. Ideally, a covariate explains variability in the dependent variable (dv) that the factors do not explain. This reduces the error variance. Reducing error variance causes F and the noncentrality parameter (NCP) (λ) for factors to become larger because the denominator of the test gets smaller. As F and λ increase, power increases. Well-chosen covariates do wonders for power. As with other designs, larger sample size, more liberal alpha error criteria, larger differences between means, and smaller standard deviations yield more power.

The value of a covariate is limited to the extent that it unrelated to the factors. This issue is similar to multicollinearity considerations discussed in Chapter 8. If factors relate to the covariate, then the covariate explains some of the variability in the dv that the factors would otherwise explain. This causes a reduction in the *F* statistic for the factors of interest. Poorly selected covariates reduce power by removing variance explained by factors and reducing error degrees of freedom.

An important assumption of ANCOVA is that the covariate and dv demonstrate the same relationship across every level of the iv. This is equivalent to assuming that there is no interaction between the covariate and the factor(s). This assumption is stated formally as the homogeneity of covariance or homogeneity of regression assumption. If you expect a covariate by factor interaction, then do not use ANCOVA. Regression approaches described in this chapter handle violations of this assumption nicely.

Although, power analyses assume a priori covariate selection, I want to further stress the importance of clearly justifying covariates in designs. Highly influential work (e.g., Simmons, Nelson, & Simonsohn, 2011) highlights how post hoc covariate inclusion increases Type I error rates. For this reason, some editors and reviewers scrutinize covariate analyses more critically than in the past.

Example 9.1: ANCOVA

In Chapter 5, an example focused on a two factor between subjects ANOVA design involving prediction of attitudes toward specific affirmative action (AA) policies based on policy type (recruitment vs. tiebreaker) and justification (none vs. increased diversity) for the policy. That design required a sample of over 1000 participants to produce power of roughly 80% for the test of the interaction. The next example examines how adding a covariate (general AA attitudes) reduces the sample size requirements.

Previous work found correlations between general AA attitudes and attitudes toward specific policies of around .40 for several applications of AA (e.g., Aberson, 2007). That is, how people feel about AA in general (e.g., "I support Affirmative Action") relates to their evaluations of specific AA policies, regardless of policy content. Based on this information, the example uses an estimated correlation between general attitudes and attitudes toward each policy of .40. Like the process involved in addressing correlations between control variables in regression analysis, this estimate should focus on accuracy rather than meaningfulness (i.e., what we expect the value to be in the population rather than how large the effect would be to have practical importance).

The anc function in Table 9.1 adds a covariate to the analyses found in Chapter 5. The format of the function follows and differs only slightly from functions used for factorial ANOVA.

anc(m1.1, m2.1, m1.2, m2.2, m1.3, m2.3, m1.4, m2.4,

s1.1, s2.1, s1.2, s2.2, s1.3, s2.3, s1.4, s2.4,

r, s, alpha, factors, n)

Means are noted with m. The first number denotes level of the first factor (up to four). The second denotes level of the second factor (2 only). Users can specify 2 × 2, 3 × 3, or 4 × 2 designs. For a one factor design, use m1.1, m2.1, etc. Standard deviations are noted with *s* and follow the same conventions as means. The value *r* specifies the correlation between the covariate and the dv. Alpha defaults to .05. Factors specifies whether the design has 1 or 2 factors. Sample size, noted by n, addresses sample size per cell.

As shown in Table 9.1, a sample of 251 per cell yields power of .86 for the interaction. Comparing power from the covariate analysis to power for the original analysis found in Table 9.2 shows there is more power for detecting effects for the factors and the interaction following addition of the covariate.

To explore the how ANCOVA increases power, Tables 9.1 (ANCOVA) and 9.2 (ANOVA) also include output for each analysis that includes sums of squares. Note that the sum of squares for both factors and the interaction are the same across the two analyses. That is, the effects explain the same amount of variance in attitudes. The difference between the two analyses is that the error variance (noted as Residuals) is smaller for the covariate analysis (2427.60) than for the ANOVA without the covariate (2890.00). The addition of the covariate accounts for 462.40 sums of squares toward the explanation of the dependent measure (note that 462.40 is the difference between 2427.60 and 2890.00).

TABLE 9.1 R Code and Output for Two Factor ANCOVA ($n = 251$ per cell)

```
anc(m1.1=.85, m2.1=2.5, s1.1 = 1.7, s2.1=1,
m1.2=0.85, m2.2=2.5, s1.2=1.7, s2.2=1,
m1.3=0.0, m2.3=2.5, s1.3=1.7, s2.3=1,
m1.4=0.6, m2.4=2.5, s1.4=1.7, s2.4=1, r=0.4,
n=251, factors=2)
## [1] "Sample size per cell = 251"
## [1] "Sample size overall = 1004"
## [1] "Power IV1 = 0.9999 for eta-squared = 0.030"
## [1] "Power IV2 = 0.8613 for eta-squared = 0.009"
## [1] "Power IV1*IV2 = 0.8613 for eta-squared = 0.009"
##             Sum Sq  Df   F value    Pr(>F)
## cov         462.40    1  190.2857 < 2.2e-16***
## iv1          75.93    1   31.2455   2.933e-08***
## iv2          22.59    1    9.2962   0.002357**
## iv1:iv2      22.59    1    9.2962   0.002357**
## Residuals 2427.60   999
```

TABLE 9.2 R Code and Output for Two Factor ANOVA (for comparison)

```
anova2x2(m1.1=0.85, m1.2=0.85, m2.1=0.00, m2.2=0.60,
s1.1=1.7, s1.2=1.7, s2.1=1.7, s2.2=1.7,
n1.1=251, n1.2=251, n2.1=251, n2.2=251,
alpha=.05)
## [1] "Power IV1 = 0.9992 for eta-squared = 0.0256"
## [1] "Power IV2 = 0.7976 for eta-squared = 0.0078"
## [1] "Power IV1*IV2 = 0.7976 for eta-squared = 0.0078"
##             Sum Sq    Df  F value     Pr(>F)
## iv1          75.93     1  26.2725   3.558e-07***
## iv2          22.59     1   7.8166   0.005276**
## iv1:iv2      22.59     1   7.8166   0.005276**
## Residuals  2890.00  1000
```

Calculations using Formulae 5.1–5.3 illustrate how the covariate increases power in this analysis. The first calculations are based on the ANCOVA, using information from Table 9.1. The interaction, with the covariate included in the analysis yields $\eta^2 = .009$. The NCP λ gets larger as the effect size increases.

$$\eta^2_{partial} = \frac{SS_{effect}}{SS_{effect} + SS_{error}} = \frac{22.59}{22.59 + 2427.6} = 0.00922$$

$$f^2 = \frac{\eta^2_{partial}}{1 - \eta^2_{partial}} = \frac{0.00922}{1 - 0.00922} = 0.00931$$

$$\lambda = f^2 df_{error} = 0.00931 * 999 = 9.30$$

Next are calculations for the analysis using the values in Table 9.2. This analysis reflects the two factor ANOVA that did not include a covariate.

$$\eta^2_{partial} = \frac{SS_{effect}}{SS_{effect} + SS_{error}} = \frac{22.59}{22.59 + 2890} = 0.00776$$

$$f^2 = \frac{\eta^2_{partial}}{1 - \eta^2_{partial}} = \frac{0.00776}{1 - 0.00776} = 0.00782$$

$$\lambda = f^2 df_{error} = 0.00782 \times 1000 = 7.82$$

Comparing the calculations for ANCOVA and ANOVA shows that the inclusion of the covariate reduced the error variance, which in turn increased the effect size and the NCP. The covariate also consumed a degree of freedom (999 vs. 1000). In the present example, the degree of freedom loss does little to influence the analysis. However, with small samples that degree of freedom influences power more. Well-chosen covariates make up for the lost degrees of freedom.

TABLE 9.3 R Code and Output for Two Factor ANCOVA ($n = 213$ per cell)

```
anc(m1.1=.85, m2.1=2.5, s1.1=1.7, s2.1=1,
m1.2=0.85, m2.2=2.5, s1.2=1.7, s2.2=1,
m1.3=.0, m2.3=2.5, s1.3=1.7, s2.3=1,
m1.4=0.6, m2.4=2.5, s1.4=1.7, s2.4=1, r=0.4,
n=213, factors=2)
## [1] "Sample size per cell = 213"
## [1] "Sample size overall = 852"
## [1] "Power IV1 = 0.9993 for eta-squared = 0.0303"
## [1] "Power IV2 = 0.7999 for eta-squared = 0.0092"
## [1] "Power IV1*IV2 = 0.7999 for eta-squared = 0.0092"
```

The next step in this analysis examines how large a sample (with the covariate included) is necessary for adequate power. Table 9.3 shows that a sample of $n = 213$ per group yields power of approximately 80% for the interaction. Although still a large sample (852 participants), there is a net savings of 152 participants over the factorial ANOVA without the covariate. At this point, a good question is whether the sample size savings justifies inclusion of the covariate measures. In this study, the covariate measure is a five-item general AA attitudes scale, requiring only about a minute of the participant's time to complete. Adding this short measure reduces the sample size requirement by 15%.

Moderated Regression Analysis (Regression with Interactions)

Moderated regression analysis focuses on regression models with interactions. Interactions can be between categorical variables, continuous variables, or both. However, ANOVA procedures handle categorical-by-categorical interactions more simply. This section includes three approaches to moderated regression analysis. Two expand on the covariate and multiple regression tests found in this chapter and in Chapter 8. These techniques work for interactions between categorical and continuous predictors as well as interactions between continuously scaled predictors. The third approach focuses on interactions between a dichotomous variable and a continuous variable. Necessary statistical values for each approach differ and are presented in the sections that address each procedure.

Factors Affecting Power

A number of issues affect power for moderated effects in regression. The first are measurement issues associated with use of continuously scaled variables. These issues include range restriction, artificial dichotomization (see Chapter 5), and poor reliability (see Aguinis, 2004). In short, if one first-order predictor (i.e., main effect)

possesses poor psychometric properties, these shortcomings also appear in the interaction term. If both first-order predictors possess poor psychometric properties, the problems are amplified for the interaction. The Additional Issues section of this chapter addresses this problem with regard to reliability. Another issue is the strength of the relationship between the first-order predictors and the criterion variable. Broadly, interaction effects are constrained by the size of these relationships. The less variance the first-order predictors explain, the smaller the possible interaction effect size (for a technical description see Rogers, 2002). The size of the interaction effect is discussed in the section that follows. As with other designs larger sample sizes and more liberal α increase power.

Size of Interaction Effects

The techniques discussed in this section focus either on estimating patterns of correlations or the amount of variance explained by the interaction. The effect size reflects the relationship between the interaction term and the dv. Many forms of interactions typically produce small effects. For example, Aiken and West note in discussing interactions between continuous variables that "[o]bserved effect sizes for interactions are very small, accounting for about 1% of the variance in outcomes. … The social scientist is forewarned" (1991, pp. 170). Similarly, a review of 30 years of publications in applied psychology and management found the median effect size for regression interactions between categorical and continuous predictors was $f^2 = .002$ or 0.2% variance explained by the interaction on the dv (Aguinis, Beaty, Boik, & Pierce, 2005).

In considering these findings, it is important to note that Aguinis et al. (2005) concentrated on areas of investigation that tend to employ designs addressing interactions between demographic variables such as gender and a measured variable. In many of these studies, tests were exploratory (e.g., does gender moderate the predictor–dv relationship?) rather than theoretically derived. Although complimentary meta-analyses examining designs that include a manipulated variable that interacts with a measured variable are not available, a cursory examination of work in fields such as social psychology suggests that when manipulated variables are included in the interaction, larger effect sizes are common. Similarly, it is reasonable to expect that when interaction hypotheses follow from well-established theory, larger effects are likely.

Previous chapters discussed the importance of designing for "meaningful" effects. When discussing interactions, it is difficult to determine what size of effect would be meaningful but there are several useful approaches to obtaining estimates. First, focus on the sort of effects detected typically in your area of inquiry. Examination of a handful of studies presenting regression interactions can give a sense of the typical effect size for the topical area. Second, it appears that regression interactions are strongest when one variable is manipulated and predictions are theoretically supported. Interactions are weakest when neither

variable is manipulated and tests are exploratory. This should serve as a reality check for power analyses. For example, effects approaching even 1% explained variance represent a relatively large interaction for a correlational design focused on exploratory analyses of interactions between two measured variables.

Regression Analogy Approaches

One regression analogy approach is conceptually the same as the multiple regression approach for coefficients found in Chapter 8. This approach treats the interaction in the same manner as the other predictors. This strategy requires estimates of correlations between all variables in the model, including the interaction. In practice, the interaction–dv correlation is often difficult to estimate, so the discussion regarding commonly observed effect sizes is particularly relevant. This approach is flexible and accommodates designs with categorical by continuous interactions as well as continuous by continuous interactions.

A second approach is to estimate the R^2 *Change* provided by the addition of the interaction to a model that includes the other predictors. This analysis proceeds according to the R^2 *Change* analyses in Chapter 8. This approach requires particular attention to effect size estimation for the interaction and an estimate of the variance explained by all the predictors. This approach is necessary for any design where the interaction degrees of freedom exceed one.

Comparison of Correlations/Simple Slopes Analogy Approach

An alternative approach uses calculations presented by Aguinis (2004). The primary difference between this approach and the regression analogy approach is that it is limited to situations with a single dichotomous predictor and a single continuous predictor. This approach focuses on the size of the correlation (or the unstandardized regression coefficient) for the relationship between the continuous predictor and the dependent measure in each of the two groups.

The analyses presented in this section use Formula 9.1 to calculate the effect size. The primary advantage to this formula is that it provides an adjustment for heterogeneity of variance that is represented in the formula through the consideration of standard deviations across the levels of the categorical moderator. Homogeneity of variance violations are common in studies examining categorical by continuous predictor interactions (Aguinis, Petersen, & Pierce, 1999).

$$f^2_{modified} = \frac{\sum (n_j - 1)\rho_j^2 \sigma^2_{y_j} - \frac{\left(\sum (n_j - 1)\rho_j \sigma_{y_j} \sigma_{x_j}\right)^2}{\sum (n_j - 1)\sigma^2_{x_j}}}{\sum (n_j - 2)(1 - \rho_j^2)\sigma^2_{y_j}} \tag{9.1}$$

The correlation/simple slope approach may be more intuitive than the first two approaches discussed in this chapter. For this approach, we address relationships

between the continuously scaled predictor and dv for each category (i.e., estimates of correlations between the predictor and dv in Group 1 and Group 2). Hypotheses that specify relationships between the predictor and dependent measure for one condition but not another can be modeled nicely with this approach.

Moderated Regression Examples

The following examples show each of the approaches discussed applied to the same example. Many of the estimates required for moderated regression analysis are not obvious, so I devote considerable space to approaches used to derive estimates from a published article.

Ayduk, Gyurak, and Luerssen (2008) examined the moderating effects of rejection sensitivity on the relationship between social rejection and aggression. The researchers exposed participants to a manipulation wherein a potential partner either rejected or did not reject them. Participants then aggressed by allocating hot sauce to the partner after being informed that he/she disliked spicy food. This example follows the design of a study to replicate these findings. The initial step takes information from the study as a guide to expected effect sizes. Ideally, authors provide a correlation matrix and standard deviations for the dv and predictors with these values also presented for each level of the categorical predictor. Of course, a single study does not necessarily provide an accurate estimate of the population effect size.

Regarding effect sizes, the effect for condition as $d = 0.34$ and the interaction as $d = 0.53$ based on $n = 122$ (see Ayduk et al., 2008. Using Formula 9.2 for converting d to ρ, these values become .17 and .26, respectively. The authors did not report the sensitivity–aggression or condition–sensitivity relationships; however, the article suggested that these were very small relationships, so it is reasonable to estimate them with zero or near zero effect sizes. Similarly, the authors did not present correlations between the interaction term and the first-order variables. Theoretically, the covariate was not expected to relate to the condition, so estimating a correlation of zero is a reasonable approach as is using a small correlation (e.g., $\rho = .05$). Since the authors did not report these values, you may have to plug and play a bit to find values that reproduce the analyses found in the research report (alternatively, you can contact the authors). Table 9.4 presents estimates of the correlations and standard deviations.

TABLE 9.4 Descriptive Statistics for Moderated Regression Example

	Aggression	*Condition*	*Sensitivity*	*C × S*
Aggression	$\sigma = 2.72$	–	–	–
Condition	$\rho = .17$	$\sigma = 0.50$	–	–
Sensitivity	$\rho = .00$	$\rho = .00$	$\sigma = 3.25$	–
C × S Interaction	$\rho = .26$	$\rho = .05$	$\rho = .05$	$\sigma = 1.00$

$$\rho = \sqrt{\frac{d^2}{d^2 + \frac{1}{p_1 p_2}}} \tag{9.2}$$

Example 9.2: Regression Analogy (Coefficients)

The first set of analyses found in Table 9.6 uses the mrc_shortcuts function to produce analyses based on the correlations. This analysis verifies that the relationships specified corresponded to the relationships the authors reported. Of particular interest are the interaction results for which the authors reported $F(1, 118) = 8.3$. With a single *df* for the interaction, *F*-change is equivalent to t^2. For the interaction, $t(118) = 2.875$. 2.875^2 is 8.27, matching the authors results. This step is not required if you are not trying to reproduce results.

Table 9.6 presents the power analysis for coefficients using the MRC function. The format of that function is detailed in Chapter 8. A sample of $n = 122$ produces power of .81 for the interaction (noted as b3 in the output).

Another important question is whether we are comfortable using the previously reported effect size as the target effect size for power. Recall that throughout the text, I encourage designing for meaningful effects so we need to add some additional context to the discussion.

TABLE 9.5 R Shortcuts to Obtain Values for R^2 *Change* (Bolded Values Used for Calculations)

```
MRC_shortcuts(ry1=.17, ry2=-.00, ry3=.26, r12=.00, r13=.05,
r23=.05, n=122)
##
## Call:
## lm(formula = X1 ~ X2 + X3 + X4, data = pop2)
##
## Residuals:
## Min 1Q Median 3Q Max
## -2.60804 -0.53056 0.09744 0.58337 2.77310
##
## Coefficients:
##              Estimate     Std. Error  t value   Pr(>|t|)
## (Intercept)   1.507e-17   8.734e-02    0.000    1.00000
## X2            1.574e-01   8.781e-02    1.792    0.07568
## X3           -1.264e-02   8.781e-02   -0.144    0.88580
## X4            2.528e-01   8.792e-02    2.875    0.00479**
## -
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1
##
## Residual standard error: 0.9647 on 118 degrees of freedom
## Multiple R-squared: 0.09247, Adjusted R-squared: 0.0694
## F-statistic: 4.008 on 3 and 118 DF, p-value: 0.009319
```

TABLE 9.6 R Code and Output for Moderated Regression (Test of Coefficient Approach)

```
MRC(ry1=.17, ry2=-.00, ry3=.26, r12=.00, r13=.05, r23=.05,
n=122)
## [1] "Sample size is 122"
## [1] "Power R2 = 0.828"
## [1] "Power b1 = 0.428"
## [1] "Power b2 = 0.052"
## [1] "Power b3 = 0.814"
```

To calculate an effect size, take the *t*-value (2.875) for the interaction, *df* residual (118), and the model R^2 (.092) to Formula 9.3.

$$R^2_{Change} = \frac{t^2}{df_{Residual}} * (1 - R^2) \tag{9.3}$$

$$R^2_{Change} = \frac{2.875^2}{118} * (1 - 0.092) = 0.064$$

The f^2 value (.07), calculated using Formula 8.7, for the interaction is large when compared to the meta-analytic results discussed, likely because the researchers experimentally manipulated one of the variables (rejection).

$$f^2_{Change} = \frac{R^2_{Change}}{1 - R^2_{Model}} = \frac{0.064}{1 - 0.092} = 0.07$$

At this point, a conservative approach suggests designing for a smaller effect size. Keep in mind that the reported effect size may or may not be a good representation of the population effect. The population effect may be larger or smaller than observed in any single study. A conservative approach when replicating studies is to reduce the effect size (f^2) to design for a study that was sensitive to the detection of smaller effects.

Example 9.3: Regression Analogy (R^2 *Change*)

The regintR2 function handles R^2 *Change* approaches. The example that follows reduces the effect size (f^2 *change*) by cutting it in half to .035. Rearranging the f^2 *change* formula yields R^2 *Change* of .032. Note that this calculation used the R^2 *Model* estimate of .092 from Table 9.5. Although this value would likely be smaller if the R^2 *Change* were smaller, using the larger estimate is more conservative.

$$R^2_{Change} = f^2_{Change}\left(1 - R^2_{Model}\right) = 0.035 * \left(1 - 0.092\right) = 0.032$$

The regintR2 function in Table 9.7 addresses power across a range of sample sizes. The format of the function is as follow:

TABLE 9.7 R Code and Output for R^2 *Change* Analysis for Interaction

```
regintR2(R2Mod=.092, R2Ch=.032, mod_pred=3, ch_pred=1,
nlow=100, nhigh=400, by=20)
[1] "Power with n = 100 = 0.4448"
[1] "Power with n = 120 = 0.5181"
[1] "Power with n = 140 = 0.5847"
[1] "Power with n = 160 = 0.6444"
[1] "Power with n = 180 = 0.6974"
[1] "Power with n = 200 = 0.7439"
[1] "Power with n = 220 = 0.7843"
[1] "Power with n = 240 = 0.8192"
[1] "Power with n = 260 = 0.8491"
[1] "Power with n = 280 = 0.8745"
[1] "Power with n = 300 = 0.896"
[1] "Power with n = 320 = 0.9142"
[1] "Power with n = 340 = 0.9294"
[1] "Power with n = 360 = 0.9421"
[1] "Power with n = 380 = 0.9526"
[1] "Power with n = 400 = 0.9614"
```

regintR2(R2Mod, R2Ch, mod_pred, ch_pred, nlow, nhigh, by)

R2Mod is R^2 *Model* and R2Ch is R^2 *Change*. The values noted as pred are number of predictors in the full model (mod_pred) and number of predictors in the interaction (i.e., degrees of freedom for interaction; ch_pred). nlow and nhigh define the range of sample sizes. by specifies the increase in sample size from nlow (e.g., if nlow is 10 and by is 5, the code produces power for 10, 15, 20, etc. until reaching nhigh).

Table 9.7 addresses power for R^2 *Change* = .032 and R^2 *Model* = .092. With this smaller R^2 *Change* value, to attain power of .80 we require $n = 240$, almost double that in the original analysis.

Example 9.4: Comparison on Correlations/Simple Slopes

The third approach to power analysis for moderated regression involves comparing relationships across the experimental conditions. The present example requires estimates of the sensitivity–aggression correlation (or regression coefficients) for the control group and the rejection group as well as standard deviations for sensitivity and aggression for each of the conditions. This approach is intuitive as it provides a direct comparison of descriptive values (rather than the effect size for the interaction). This approach is more accurate than the others when the distribution of variances on the dv across levels of the categorical moderator are unequal (i.e., heterogeneity of variance).

The article (Ayduk et al., 2008) reported the standard deviations for aggression as 3.22 and 2.10 for the rejection and control conditions, respectively. However, the authors reported the standard deviation for the sensitivity score for only the entire sample (3.25). This may seem problematic, but unless there is an expectation of heterogeneity of variances, we can use the same standard deviation for both groups. The authors reported unstandardized regression coefficients (b) for the sensitivity–aggression relationship in the control condition as $b = -0.17$ and as $b = 0.25$ for the rejected condition. Using Formula 9.4 to convert these values finds correlations of –.26 and .25. Table 9.8 summarizes the descriptive statistics used in the analysis.

$$\rho = b\frac{\sigma_x}{\sigma_y}$$

$$\rho_{control} = -0.17\frac{3.25}{2.10} = -0.26 \tag{9.4}$$

$$\rho_{rejected} = 0.25\frac{3.25}{3.22} = 0.25$$

The key to this analysis is the calculation of the f^2 statistic from the group-based statistics. An example of this calculation, using a sample of $n = 61$ per group appears below. Given $\lambda = 7.92$ with $df = 1, 118$ and $\alpha = .05$, power is .797. This result is consistent with the earlier analysis using coefficients.

$$f^2_{modified} = \frac{\sum(n_j - 1)\rho_j^2\sigma_{y_j}^2 - \frac{\left(\sum(n_j - 1)\rho_j\sigma_{y_j}\sigma_{x_j}\right)^2}{\sum(n_j - 1)\sigma_{x_j}^2}}{\sum\left(n_j - 2\right)\left(1 - \rho_j^2\right)\sigma_{y_j}^2}$$

$$= \frac{\left[\left(\left(60 * -0.26^2 * 2.10^2\right) + \left(60 * 0.25^2 * 3.22^2\right)\right) - \frac{\left[\left(60 * -0.26 * 2.10 * 3.25\right) + \left(60 * 0.25 * 3.22 * 3.25\right)\right]^2}{\left(60 * 3.25^2\right) + \left(60 * 3.25^2\right)}\right]}{\left(59 * \left(1 - 0.26^2\right) * 2.10^2\right) + \left(59 * \left(1 - 0.25^2\right) * 3.22^2\right)}$$

$$= 0.0671$$

$$\lambda = f^2 df_{error} = 0.0671 * 118 = 7.92$$

TABLE 9.8 Descriptive Statistics by Group for Moderated Regression

	Control	*Rejected*
σ_x	3.25	3.25
σ_y	2.10	3.22
ρ	–.26	.25

TABLE 9.9 R Code and Output for Group-based Interaction Tests

```
regint(Group1=-.26, Group2=.25, alpha=.05, Prop_n1=0.5,
nlow=110, nhigh=140, by=2, Estimates=1)
[1] "Power with n1 = 55 n2 = 55 = 0.7751"
[1] "Power with n1 = 56 n2 = 56 = 0.7827"
[1] "Power with n1 = 57 n2 = 57 = 0.79"
[1] "Power with n1 = 58 n2 = 58 = 0.7972"
[1] "Power with n1 = 59 n2 = 59 = 0.8041"
[1] "Power with n1 = 60 n2 = 60 = 0.8108"
[1] "Power with n1 = 61 n2 = 61 = 0.8174"
[1] "Power with n1 = 62 n2 = 62 = 0.8237"
[1] "Power with n1 = 63 n2 = 63 = 0.8298"
[1] "Power with n1 = 64 n2 = 64 = 0.8358"
[1] "Power with n1 = 65 n2 = 65 = 0.8416"
[1] "Power with n1 = 66 n2 = 66 = 0.8472"
[1] "Power with n1 = 67 n2 = 67 = 0.8526"
[1] "Power with n1 = 68 n2 = 68 = 0.8578"
[1] "Power with n1 = 69 n2 = 69 = 0.8629"
[1] "Power with n1 = 70 n2 = 70 = 0.8678
## [1] "Effect size (R2 Change/Squared Semi Partial) = 0.066"
```

Table 9.9 demonstrates uses of the regint function for power calculations. The format of the function is as follow:

> regint(Group1, Group2, Estimates, sx1, sx2, sy1, sy2, alpha, Prop_n1, nlow, nhigh, by)

The values Group1 and Group2 are either correlations or unstandardized regression coefficients for the iv–dv relationship. Estimates = 1 reflects use of correlation, Estimates = 2 is for coefficients. The sx and sy values reflect standard deviations for the continuously scaled predictor (sx1 and sx2) and dv (sy1 and sy2). Prop_n1 indicates the overall proportion of the sample in the first group. As before, nlow, nhigh, and by define the range of sample sizes.

Logistic Regression

Logistic regression (LR) involves predicting a dichotomous outcome from either categorical or continuous predictor variables. This section covers designs with a single dichotomous predictor, a single continuous predictor, and multiple predictors.

Necessary Statistics

Power analyses for LR with a categorical predictor require proportion of outcomes broken down by cell. Specifically, the proportion of people in the first

category of the predictor who had a favorable outcome on the dv (p1) and the proportion of people in the other category of the predictor with a favorable outcome (p0). Additionally, some calculations require the proportion of the total sample expected in the first category of the predictor (prop), event rate (ER), and how well one predictor is explained by the others in the model (R2).

Expanding on an example appearing in Cohen et al. (2003), Examples 9.5–9.7 examine power for LR models with a single dichotomous predictor, a single categorical predictor, and multiple predictors.

Example 9.5: Logistic Regression with a Single Categorical Predictor

This example examines predicting compliance with mammography screening recommendations (in compliance vs. not in compliance). The predictor of compliance is doctor's recommendations (received recommendation vs. did not receive recommendation). Table 9.10 provides descriptive statistics for a sample of 164 women. I use these values to estimate values for a replication study with Power = .95. The table provides examples of the calculation of the proportion in the recommended group who complied [p(1)] and those in the no recommendation group who complied [p(0)]. The value prop reflects the proportion of the sample in the recommendation group.

Formula 9.5 provides an estimate of sample size for a dichotomous predictor (Hseih, Bloch, & Larsen, 1998).

$$N = \frac{\left(z_{1-alpha/2}\sqrt{\frac{ER(1-ER)}{prop}} + z_{1-beta}\sqrt{p_0(1-p_0) + \frac{p_1(1-p_1)(1-prop)}{prop}} \right)^2}{(p_0 - p_1)^2(1-prop)} \tag{9.5}$$

Where ER is the event rate. The ER corresponds to the overall proportion with a favorable outcome. The z-values reflect the z-score corresponding each proportion. For example, $z_{1-\beta}$ for Power = .95 is 1.64. This is the z-score at the 95%ile.

TABLE 9.10 Descriptive Statistics for One Categorical Predictor

	Recommend Yes	*Recommend No*
Complied	69	7
Did not Comply	44	44
	$p(1) = 69/(69+44) = .611$	$p(0) = 7/(7+44) = .137$
	prop = 69 + 44/(69 + 4 + 7 + 44) = .689	–
	$\overline{p} = 76/164 = .463$	–
Odd comply	69/44 = 1.568	7/44 = 0.159

TABLE 9.11 R Code and Output for Logistic Regression, One Categorical Predictor

```
LRcat(p0=.137, p1=.611, prop=.689, power=.95)
## [1] "Sample Size = 55 for Odds Ratio = 9.894"
```

Applying the formula to the example, using $\alpha = .05$ and Power = .95 yields a sample size of 55.

$$N = \frac{\left(1.96\sqrt{\frac{0.463(1-0.463)}{0.689}} + 1.64\sqrt{0.137(1-0.137)+\frac{0.611(1-0.611)(1-0.689)}{0.689}}\right)^2}{(0.137-0.611)^2(1-0.689)}$$

$$= \frac{\left(1.96\sqrt{\frac{0.2486}{0.689}} + 1.64\sqrt{0.1182+\frac{0.0739}{0.689}}\right)^2}{0.2247 * 0.3110}$$

$$= \frac{\left(1.96(0.6007)+1.64(0.4749)\right)^2}{0.0699}$$

$$= \frac{1.9562^2}{0.0699}$$

$$= 54.7$$

Table 9.11 demonstrates us of the LRcat function to complete these calculations. The format of the function is:

LRcat(p0, p1, prop, alpha, power)

The values for p0 and p1 are the probability of a desirable outcome in the control and treatment conditions, respectively. Prop is the proportion in the treatment condition. Alpha defaults to .05.

As shown in Table 9.11, a sample of 55 provides Power = .95. The output also provides an odds ratio. The odds ratio is odds for the treatment group over the odds for the control group (1.568/.159 = 9.86).

Example 9.6: Logistic Regression with a Single Continuous Predictor

This example examines predicting compliance with mammography screening recommendations (in compliance vs. not in compliance) from perceived benefits of mammography. Table 9.12 provides descriptive statistics for this relationship as well as several others (used in Example 9.7). In the present example, there is not an odds ratio (OR) provided, however, the correlation of .36 converts to an OR using Formula 9.6

TABLE 9.12 Descriptive Statistics for Logistic Power Examples

	Comply	*Recommend*	*Know*	*Benefits*	*Barriers*
Comply	–	–	–	–	–
Recommend	.44	–	–	–	–
Knowledge	–.06	–.08	–	–	–
Benefits	.36	.40	–.01	–	–
Barriers	–.41	–.31	.06	–.39	–
Prop Yes (M)	.46	.69	0.62	4.16	1.42
SD	n/a	n/a	0.18	1.04	1.38

$$
\begin{aligned}
OR &= \exp\frac{\left(\left(2r/\sqrt{(1-r^2)}\right)*\pi\right)}{\sqrt{3}} \\
OR &= \exp\frac{\left(\left(2*0.36/\sqrt{(1-0.36^2)}\right)*\pi\right)}{\sqrt{3}} \\
&= \exp\frac{\left(\left(0.72/\sqrt{(0.8704)}\right)*3.1416\right)}{\sqrt{3}} \\
&= \exp\frac{(0.7717*3.1416)}{\sqrt{3}} \\
&= \exp\frac{2.4245}{\sqrt{3}} \\
&= 4.05
\end{aligned}
\tag{9.6}
$$

Formula 9.7 provides the calculation for sample size.

$$
\begin{aligned}
N &= \frac{\left(z_{1-alpha/2}+z_{1-beta}\right)^2}{ER*(1-ER)*\lg(OR)^2} \\
N &= \frac{\left(1.96+1.64\right)^2}{0.463*(1-0.463)*\lg(4.05)^2} \\
&= \frac{12.96}{0.463*(0.537)*1.956} \\
&= \frac{12.96}{0.463*(0.564)*1.956} \\
&= \frac{12.96}{0.4863} \\
&= 26.67
\end{aligned}
\tag{9.7}
$$

Table 9.13 demonstrates use of the LRCont function. The form of the function is:

LRcont(OR, ER, r, power, alpha, R2)

TABLE 9.13 R Code and Output for Logistic Regression, One Continuous Predictor

```
LRcont(OR=4.05, ER=.463, power=.95)
## [1] "Sample Size = 27, Odds Ratio = 4.05"
```

OR is the odds ratio. Instead of OR, the function accepts entry of the correlation (r) directly. ER is the event rate (estimated from 76 complied over total sample size of 164). Power defaults to .80 and alpha defaults to .05. R2 is discussed in the context of Example 9.7. It is set to a default of .00.

Example 9.7: Power for One Predictor in a Design with Multiple Predictors

This example addresses power for one predictor within a model that contains additional predictor. Specifically, the power for Benefits in a model that includes Knowledge, Barriers, and Recommendation. This approach requires a single piece of additional information, how well the other variables in the model explain the predictor of interest (termed R^2 in Formula 9.8).

To calculate R^2, we can use the shortcut code from Chapter 8. The code will handle up to four predictors. In this context, the predictor of interest serves as the dependent measure. Table 9.14 finds $R^2 = .239$. Correlations come from Table 9.12

TABLE 9.14 R Code and Output for R^2 Estimation

```
MRC_shortcuts(ry1=.40, ry2=-.01, ry3=-.39, r12=-.08, r13=-.31,
r23=.06, n=164)
##
## Call:
## lm(formula = X1 ~ X2 + X3 + X4, data = pop2)
##
## Residuals:
## Min 1Q Median 3Q Max
## -2.81913 -0.49634 0.03331 0.50482 2.17843
##
## Coefficients:
##              Estimate    Std. Error   t value    Pr(>|t|)
## (Intercept)   3.756e-17  6.874e-02     0.000     1.000
## X2            3.110e-01  7.268e-02     4.279     3.22e-05***
## X3            3.261e-02  6.922e-02     0.471     0.638
## X4           -2.956e-01  7.257e-02    -4.072     7.30e-05***
## -
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1
##
## Residual standard error: 0.8803 on 160 degrees of freedom
## Multiple R-squared: 0.2393, Adjusted R-squared: 0.2251
## F-statistic: 16.78 on 3 and 160 DF, p-value: 1.582e-09
```

TABLE 9.15 R Code and Output for One Continuous Predictor with Other Variables in Model

```
LRcont(r=.36, ER=.463, power=.95, R2=.239)
## [1] "Sample Size = 36, Odds Ratio = 4.0543"
```

Formula 9.8 demonstrates an adjustment to the sample size based on the analysis in Table 9.13 (see Hsieh et al., 1998).

$$N_{new} = \frac{N}{1-R^2}$$
$$N_{new} = \frac{27}{1-0.239} = 35.5 \tag{9.8}$$

Table 9.15 demonstrates use of the LRcont to complete the calculation.

Mediation (Indirect Effects)

Mediated or indirect effects address how well an intervening variable explains the relationship between a predictor variable and an outcome. There are many possible mediation models. This section provides power analyses for single and multiple parallel mediation models. For more information on mediation, Hayes (2017) provides details regarding a wide range of models. I use the terms "mediator" and "indirect" effects interchangeably in this section. Often mediation implies a causal effect, so I prefer the term indirect effects but use both since mediator is more commonly used.

Factors Affecting Power

Figure 9.1 provides a basic overview of values that influence on power. The primary values of interest are the "*a*" and "*b*" paths. The a path represents the regression coefficient of a model predicting the mediating variable from the independent variable (*m* predicted by *x*). Conceptually, this is simply a function of the strength of the *x*-*m* correlation. The b path is the regression coefficient

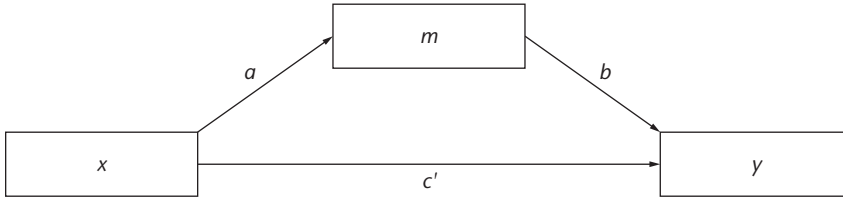

FIGURE 9.1 Model of Effects in Mediation Analysis.

TABLE 9.16 Power for Indirect Effects by Size of Relationship ($n = 100$)

	$r_{xy} = .10$		
	$r_{my} = .10$	$r_{my} = .30$	$r_{my} = .50$
$r_{mx} = .10$	.10	.16	.17
$r_{mx} = .30$	.11	.58	.79
$r_{mx} = .50$	.08	.77	.99
	$r_{xy} = .30$		
	$r_{my} = .10$	$r_{my} = .30$	$r_{my} = .50$
$r_{mx} = .10$	.09	.16	.17
$r_{mx} = .30$	.03	.47	.76
$r_{mx} = .50$	.09	.42	.95
	$r_{xy} = .50$		
	$r_{my} = .10$	$r_{my} = .30$	$r_{my} = .50$
$r_{mx} = .10$	.07	.16	.17
$r_{mx} = .30$	.08	.36	.74
$r_{mx} = .50$	.49	.10	.85

for a model predicting *y* from both *x* and *m*. The size of the b path is determined by both the correlation between *m* and *y* and the *x*-*m* correlation (see Chapter 8 for calculation examples). The indirect (also know as "mediated") effect is *a* times *b* (*ab*). As *ab* rises, power increases.

Table 9.16 explores power (using $n = 100$) for indirect effects across a variety of situations. Correlations between the predictor and mediator, the predictor and outcome variable, and the mediator and outcome variable are presented in varying sizes (.1, .3, and .5). For most situations, stronger correlations between the mediator and outcome variable produce more power for indirect effects.

Necessary Statistics

Correlations between all variables in the model are the only required values for addressing power. This is identical to multiple regression models.

Example 9.8: One Mediating Variable

This example examines power for a single indirect effect. Table 9.17 presents correlations for several relationships. This section examines the role of anxiety in mediating the relationship between contact and attitudes. Table 9.18 uses the

TABLE 9.17 Correlations for Indirect Effects Examples

	Neg. Contact	*Anxiety*	*Real Threat*	*Symb. Threat*
Neg. Contact	–	–	–	–
Anxiety	.25	–	–	–
Real Threat	.30	.40	–	–
Symb. Threat	.30	.40	.70	–
Attitudes	–.35	–.50	–.50	–.50

TABLE 9.18 R Code and Output for Indirect Effects with a Single Mediator

```
med(rxm1=.25, rxy=-.35, rym1=-.5, mvars=1, n=150)
## [1] "Power for n = 150 mediator 1 = 0.8057"
```

med function to calculate power for a sample of 150. The structure of the function is as follows:

med(rxm1, rxy, rym1, mvars, n, alpha)

rxm1 is the correlation between the predictor variable (*x*) and the mediator (m1). rxy is the correlation between the predictor variable (*x*) and the outcome (*y*). rym1 is the correlation between the outcome variable (*y*) and the mediator (m1). mvars is the number of mediating variables. The function allows up to four. *n* is sample size and alpha defaults to .05.

Example 9.9: Multiple Mediating Variables

The example expands Example 9.8 by adding realistic and symbolic threat as additional mediating variables. Table 9.19 demonstrates use of the med function

TABLE 9.19 R Code and Output for Indirect Effects with Multiple Mediators ($n = 150, 335$)

```
med(rxm1=.3, rxm2=.3, rxm3=.25, rxy=-.35, rym1=-.5, rym2=-.5,
rym3=-.5, rm1m2=.7, rm1m3=.4, rm2m3=.4, mvars=3, n=150)
## [1] "Power for n = 150 mediator 1 = 0.4675"
## [1] "Power for n = 150 mediator 2 = 0.4675"
## [1] "Power for n = 150 mediator 3 = 0.7228"
## [1] "Power for n = 150 Total Mediation = 0.9757"
med(rxm1=.3, rxm2=.3, rxm3=.25, rxy=-.35, rym1=-.5, rym2=-.5,
rym3=-.5, rm1m2=.7, rm1m3=.4, rm2m3=.4, mvars=3, n=335)
## [1] "Power for n = 335 mediator 1 = 0.8016"
## [1] "Power for n = 335 mediator 2 = 0.8016"
## [1] "Power for n = 335 mediator 3 = 0.968"
## [1] "Power for n = 335 Total Mediation = 1"
```

for calculation of power. The function uses m1, m2, and m3 to define each mediator. The output shows that with multiple mediators, a substantially larger sample is required to reach Power = .80. Of note is that the first two mediators (realistic threat and symbolic threat) had substantially less power than the third mediator (anxiety). The two threat variables correlate strongly, reducing the size of the indirect effect substantially. Multicolinearity reduces power for indirect effect in the same manner as for multiple regression.

Additional Issues

Reliability for Interactions

Chapter 8 included a discussion of how reliability affects power for regression analyses. Interactions between continuously scaled predictors complicate this problem. The reliability of the interaction is a product of the reliabilities of the interacting variables. Practically this means the reliability for the interaction is usually lower than reliability for first-order effects (i.e., main effects). Formula 9.9 details the calculation for the reliability of an interaction. This calculation is accurate only for centered predictors (Aiken & West, 1991). Formula 9.10 calculates the observed effect size for the interaction after adjusting the true effect size for reliability of the measures.

$$\alpha_{1x2} = \frac{\rho_{12}^2 + \alpha_1\alpha_2}{\rho_{12}^2 + 1} \tag{9.9}$$

$$\rho_{obs} = \rho_{true}\sqrt{\alpha_{1x2}\alpha_y} \tag{9.10}$$

The first calculation example reflects a design with two continuously scaled predictors; each demonstrates strong reliability (α_1 and $\alpha_2 = .90$) as does the dependent measure ($\alpha_y = .90$). The relationship between the predictors in the population is small ($\rho_{12} = .20$) and the relationship between the interaction and the dependent measures in the population is $\rho_{true} = .10$. The value α_{1x2} represents reliability for the interaction. Even with measures demonstrating considerable reliability, the interaction reliability is lower than for the first-order effects ($\alpha_{1x2} = .82$).

$$\alpha_{1x2} = \frac{0.20^2 + 0.90 * 0.90}{0.20^2 + 1} = \frac{0.85}{1.04} = 0.82$$

The interaction effect observed, as compared to the relationship in the population, is related to the reliability of both the interaction and the dv. Taking these values to Formula 9.5, the correlation observed in the sample is .086 whereas the population correlation is .100. This may seem like a small reduction (.100 to .086) but it does reflect a 14% drop in effect size.

$$\rho_{obs} = 0.1\sqrt{0.82 * 0.90} = 0.086$$

Now consider a situation where the reliabilities are more modest (α_1, α_2, and $\alpha_y = .80$). This produces a 28% reduction in the expected observed effect (from .100 in the population to .072 in the sample). Most researchers would not consider $\alpha = .80$ poor reliability, however when dealing with regression interactions, even moderate departures from perfect reliability produce considerable reductions in observed effect sizes.

$$\alpha_{1x2} = \frac{0.20^2 + 0.80 * 0.80}{0.20^2 + 1} = 0.65$$

$$\rho_{obs} = 0.1\sqrt{0.65 * 0.80} = 0.072$$

Summary

This chapter addressed power analysis for ANCOVA, regression interactions, LR, and mediated effects. ANCOVA requires estimates of correlations between variables as well as patterns of mean differences. For covariate designs, a particular concern is selection of covariates that relate to the dependent measure but are unrelated to the factors. For regression interactions, several approaches exist, with all requiring some estimate of the size of the relationship between the interaction term and the dependent measure. Power for LR requires an estimate of the odds ratio (or related) values, event rates, proportional group sizes, and multicollinearity between predictors. Mediation power calculations requires correlations between all variables in the model.

10

PRECISION ANALYSIS FOR CONFIDENCE INTERVALS

Precision analyses (also known as accuracy in parameter estimation) focus on the width of confidence intervals (CIs). Precision analysis provides information that supplements power analyses and in some cases is more appropriate to research goals. Power analysis determines the likelihood of rejecting a null hypothesis given a particular population effect size, sample size, and Type I error rate. However, rejecting the null hypothesis is only half of the story. Another important issue is what range of values is reasonable to expect for the population given the sample result. A CI provides this information but can be very wide or very narrow. The wider the confidence limits, the less precise the results are. We can design for more precise (i.e., narrower) CIs around effect sizes or raw values (e.g., mean differences). However, increasing precision requires larger samples or a better design to reduce error variability.

Power analyses and precision analyses often reflect different research goals. For example, if a researcher compared two established HIV risk reduction interventions (e.g., psychoeducational interventions and cognitive–behavioral approaches) the primary question of interest would likely be whether the treatments are differentially effective. This question fits nicely with power analysis. In the design phase, the researcher determines how large differences would have to be to be practically meaningful and then decides on an appropriate sample size using the power analysis techniques discussed in previous chapters.

Now consider a project addressing how much a cognitive–behavioral intervention reduces HIV risk over no intervention. In this study, it would be hard to imagine that an established intervention based on sound psychological theory would not reduce risk. Instead, a better question is how much the technique reduces risk. For this question, power analysis would address the sample size necessary to support claims of a nonzero effect. More relevant is how large the

effect is and what the effect might reasonably look like in the population. For example, if we wanted to estimate the population effect within 0.20 units of standard deviation, precision analyses would establish the sample size necessary to produce this estimate.

Necessary Information

This chapter covers precision analyses for confidence limits around mean differences, correlations, and confidence limits based on noncentral distributions for effect sizes such as Cohen's *d* for mean differences and R^2 *Model*. For all analyses, the primary information is the desired width of the CI (see the section on determining levels of precision). Tests involving mean differences require means, standard deviations, and proportional sample sizes (what proportion in Group 1, what proportion in Group 2) for each group. For correlations, only ρ is necessary. Confidence limits on effect sizes such as *d* or R^2, involve the effect size and, if relevant, degrees of freedom.

Confidence Intervals

Before discussing precision analysis, it is useful to review CIs. Many sources argue that CIs are superior to traditional null hypothesis significance testing procedures (see Finch, Thomason, & Cumming, 2002; Hunter, 1997; and Nickerson, 2000; see also Belia, Fidler, Williams, & Cumming, 2005 and Cumming & Finch, 2005 for a discussion of misunderstandings of CIs). Whereas null hypothesis significance tests yield a simple dichotomy of outcomes (reject or fail to reject), confidence limits provide more information and better quality of information. For instance, CIs indicate a reasonable range of values for a parameter, with values outside of the confidence limits being relatively implausible. In addition, the distance between the upper and lower limits of the CI indicates the precision of the result. Finally, confidence limits allow for the same decisions about the null hypothesis as significance testing procedures do. Hypothesized values that fall outside of the confidence limits allow for rejection of a null hypothesis at a probability corresponding to the CI (e.g., values falling outside of a 95% CI correspond to $p < .05$ to; values outside a 90% CI indicate $p < .10$).

One reason CIs are valuable is because CIs provide information that is not clearly provided by other statistical values. Imagine the following situations:

> Situation 1: Group 1 ($M = 5.0$; $SD = 1.4$) outperformed Group 2 ($M = 3.2$; $SD = 1.5$), $t(10) = 2.23$, $p = .025$, $d = 1.29$.

> Situation 2: Group 1 ($M = 5.0$; $SD = 1.4$) outperformed Group 2 ($M = 3.2$; $SD = 1.5$), 95% CI around mean difference [0.0003, 3.700], $d = 1.29$.

The first situation shows a statistically significant effect and a large effect size. The second situation represents the same differences between the groups and a CI that suggests plausible values for the mean differences in the population are somewhere between large (3.7 points) and miniscule (0.0003 points). Both examples reflect the same data but the CI presentation clearly suggests limited confidence regarding the size of the differences between the two groups. A narrower CI (e.g., ranging from 1.5 to 2.1) supports a stronger conclusion about how much the groups likely differ in the population. Precision analysis allows for determination of sample size requirements that produce confidence limits of a desired width.

Types of Confidence Intervals

Interval estimates around mean differences are included in most statistical packages. Constructing this sort of interval requires taking the differences between two sample means plus or minus margin of error. For example, for the CI around the difference between two means, the margin of error involves a *t*-statistic corresponding to the confidence level multiplied by an index of standard error. Intervals of this type are often termed central intervals as their calculation involves the central *t*-distribution.

Less commonly presented are interval estimates around effect sizes (e.g., Thompson, 2002). Use of such values fits nicely with recommendations to present effect sizes and confidence limits (e.g., Wilkinson & Task Force on Statistical Inference, 1999), so presentation of these values to should become increasingly common (although they have not in the period between the first edition of the book and the present edition). A CI around an effect size yields information about likely values for the effect size in the population. This concept is appealing as it opens the door to determining what effect sizes are likely or unlikely for the population. For example, a CI of 0.40 to 0.80 drawn around an observed effect size *d* suggests that it would be unlikely for the standardized difference between means in the population to be smaller than 0.40.

CIs around effect sizes require specialized calculations because these intervals involve noncentral distributions that require iterative procedures to achieve accurate calculations. In short, there is no simple approach for deriving CIs for noncentral distributions by hand. A full explanation of the calculation of these intervals is outside of the scope of this book. Both Smithson (2003) and Kelley and Rausch (2006) provide calculation details.

Example 10.1: Confidence Limits around Differences between Means

For independent group comparisons, Formula 10.1 presents the 95% CI around the difference between two means. The right hand side of the formula (following the ± symbol) defines the precision. This is commonly termed the margin

of error. Sample size exerts considerable influence over the standard error of the differences between means. As sample size rises, the standard error decreases, making for a smaller margin of error. Larger samples therefore give results that are more precise.

Formula 10.1 notes a 95% CI. For other CIs (e.g., 99%), simply replace $t_{.95}$ with the *t*-value corresponding to the appropriate interval. Regarding notation, I use $t_{.95,\ 2\text{-}tailed}$ to represent the *t*-value in Formula 10.1. This value corresponds to the two-tailed critical value for *t* with $\alpha = .05$. Other sources might note this value as $t_{.975}$, indicating the *t*-value where 97.5% of the distribution falls at or below or $t_{.05}$.

$$(\bar{x}_1 - \bar{x}_2) \pm t_{0.95,2-tailed} s_{\bar{x}_1 - \bar{x}_2} \tag{10.1}$$

In Chapter 3, one example examined a study designed to detect a difference between exam score means of 2 points (corresponding to $d = 0.40$) when comparing students who completed a computer tutorial to those completing a standard laboratory assignment. In that example, a sample of $n = 99$ per group (198 overall) was necessary to achieve power of .80 to detect a 2-point mean differences between the groups. Imagine that for this study, we instead wanted to make particular claims regarding the how much improvement could reasonably be expected in the population (i.e., an estimate of the true effect).

The md_prec function demonstrated in Table 10.1 creates a series of CIs based on the mean difference ($u_1 - u_2 = 2.0$) shown in the Chapter 3 example. The format of the function is as follows:

md_prec(m1, m2, s1, s2, nlow, nhigh, propn1, ci, by)

The values *m* and *s* values reflect the means and standard deviations of each group. nlow and nhigh define the range of sample sizes. by specifies the increase in sample size from nlow. propn1 defines the proportion in the first group. This defaults to .50 for equally sized groups. ci determines the type of CI.

Table 10.1 demonstrates use of the function and provides output for a 2-point difference. This was the minimum difference between means termed meaningful for that Example 3.1. For precision analysis, the difference between the means does not influence precision. Regardless of the difference between means in the sample, given that the expected population standard deviations are accurate, the precision of the interval remains the same.

In thinking about precision, a good place to begin is consideration of the standard deviation. The present example involves a measure with an expected standard deviation of 5.0. Thinking about these values as they relate to the mean difference of 2.0 provides additional context. If our sample mean differed by 2.0, an interval width of 4.0 would correspond to a 95% CI around $u_1 - u_2$ that ranged from 0.0 to 4.0. An interval width of 2.0 would produce a 95% CI around $u_1 - u_2$ that ranged from 1.0 to 3.0. If our standard deviations were 20.0,

analyses producing these intervals would suggest greater precision. For example, with a standard deviation of 5.0, a width of 4 points would be large in comparison to a width of 4.0 when the standard deviation is 20.0.

Before discussing the analysis, it is important to review what power tells us and what confidence limits tell us in the context of this example. In Chapter 3, we determined that if the true population difference was 2 points (i.e., the tutorial improved scores by 2 points) that 80% of the samples drawn from this population with $n = 99$ per group would allow for rejection of the null hypothesis. Confidence limits drawn around our sample provide different information. Specifically, CIs indicate what sort of population mean differences might reasonably produce the differences observed in the sample. Even with adequate power to detect a 2-point difference, we may not be able to conclude that a difference of less than 2 points is unlikely in the population. For example, if the sample means differed by 3.5 points (considerably larger than what we termed a meaningful difference) a 95% CI with an interval width of 4.0 would range from 1.5 to 5.5. Although the CI clearly rules out conclusions of no difference between the means ($u_1 - u_2 = 0$) it does not rule out differences that are smaller than meaningful ($1.5 \leq (u_1 - u_2) < 2.0$).

As shown in Table 10.1, 100 participants per group and a mean difference of 2 points found in the sample, the CI for the population difference would range from 0.6 to 3.4. Reasonable estimates of the actual difference between the population means includes not only some small differences (e.g., 0.6) but also some large ones (e.g., 3.4). The table provides the sample size for each group (n1, n2), the lower limit (LL) and upper limit (UL) of the CI, and the precision of the interval which is simply the range between the UL and the LL. Please note that you cannot accurately convert the CI presented in this table to effect size intervals by dividing the mean values by the pooled standard deviation (I discuss this issue in more detail in the section titled "Confidence Intervals around Effect Sizes).

Of note here is that in Example 3.1, power of .80 corresponded to a sample size of 99 per group. Table 10.1 suggests that 100 per group produces confidence limits from roughly 0.60 to 3.40. Even with Power = .80, precision remains low.

Determining Levels of Precision

Unlike power analysis where power of .80 (or .95) is often considered a standard; there is no de facto standard for precision analysis. The desired level of precision can be expected to vary widely across applications but a primary issue to consider for all situations is the consequence of a lack of precision. For example, a study of the absorption of a drug likely requires considerable precision as this information lends itself to dosage decisions. However, for most behavioral science fields, decisions regarding determination of an adequate level of precision are less clear. In the previous example, imagine we designed for a CI width of 2.0, and the sample results indicated that the true difference between the computer tutorial and standard laboratory was between 1 and 3

TABLE 10.1 R Code and Output for Confidence Interval around Mean Differences Precision Analysis

```
md_prec(m1=2, m2=0, s1=5, s2=5, nlow=100, nhigh=1600, propn1=.5,
ci=.95, by=100)
## [1] "n1 = 50, n2 = 50, d = 0.4, LL = 0.0152, UL = 3.9748,
precision = 3.9596"
## [1] "n1 = 100, n2 = 100, d = 0.4, LL = 0.5977, UL = 3.3973,
precision = 2.7996"
## [1] "n1 = 150, n2 = 150, d = 0.4, LL = 0.8554, UL = 3.1413,
precision = 2.2859"
## [1] "n1 = 200, n2 = 200, d = 0.4, LL = 1.009, UL = 2.9885,
precision = 1.9795"
## [1] "n1 = 250, n2 = 250, d = 0.4, LL = 1.1137, UL = 2.8843,
precision = 1.7706"
## [1] "n1 = 300, n2 = 300, d = 0.4, LL = 1.191, UL = 2.8073,
precision = 1.6163"
## [1] "n1 = 350, n2 = 350, d = 0.4, LL = 1.2511, UL = 2.7475,
precision = 1.4964"
## [1] "n1 = 400, n2 = 400, d = 0.4, LL = 1.2995, UL = 2.6992,
precision = 1.3997"
## [1] "n1 = 450, n2 = 450, d = 0.4, LL = 1.3396, UL = 2.6593,
precision = 1.3197"
## [1] "n1 = 500, n2 = 500, d = 0.4, LL = 1.3735, UL = 2.6255,
precision = 1.252"
## [1] "n1 = 550, n2 = 550, d = 0.4, LL = 1.4027, UL = 2.5964,
precision = 1.1937"
## [1] "n1 = 600, n2 = 600, d = 0.4, LL = 1.4282, UL = 2.571,
precision = 1.1428"
## [1] "n1 = 650, n2 = 650, d = 0.4, LL = 1.4506, UL = 2.5486,
precision = 1.098"
## [1] "n1 = 700, n2 = 700, d = 0.4, LL = 1.4706, UL = 2.5287,
precision = 1.0581"
## [1] "n1 = 750, n2 = 750, d = 0.4, LL = 1.4886, UL = 2.5108,
precision = 1.0222"
## [1] "n1 = 800, n2 = 800, d = 0.4, LL = 1.5048, UL = 2.4946,
precision = 0.9898"
```

points of improvement in the population. In a worst-case scenario, we would find a difference of 1 point favoring the computer tutorial. A 1.0-point improvement is less than desirable given that implementation of the tutorial assignment involves several hours of work from instructors.

Now compare the computer tutorial situation to one involving estimates of drug absorption. Poor precision estimates could lead to absorption is less than desired. In this case, patients end up with less medicine than intended. Similarly, absorption might be greater than expected, potentially causing overdose. Clearly, considerable precision is required as the cost of an imprecise estimate may have serious health consequences for patients. In this context, the consequences of imprecision in the computer tutorial example are minor.

Confidence Intervals around Effect Sizes

Confidence limits for effect sizes provide both an index of the likely population value of the effect size and valuable information for comparing standardized values across completed studies. When I first heard the term confidence interval around an effect size, my initial thought was that calculation involved taking a regular CI, such as one around mean differences as described in the previous section, and converting it to an effect size CI by dividing the LL and UL of the mean difference by the standard deviation. This approach sometimes provides a reasonable approximation, but an accurate calculation requires far more work. Most sources simply say something to the effect of "let the computer do this." These calculations are outside the scope of the present text but Steiger and Fouladi (1997) and Kelley (2007a) offer considerable insight on the concepts and calculations. The sections that follow present computer-based approaches for each analysis.

Example 10.2: Confidence Limits around *d*

Precision analysis for *d* use the function d_prec demonstrated in in Table 10.3. This function uses calculations addressed by the MBESS package (Kelley, 2007b). The format of the function is as follows:

```
d_prec(d, nlow, nhigh, propn1, ci, by)
```

The value *d* represents the effect size expressed as a standardized mean difference. nlow and nhigh define the range of sample sizes. by specifies the increase in sample size from nlow. propn1 defines the proportion in the first group. This defaults to .50 for equally sized groups. ci determines the type of CI.

The CI for the population effect size shown in Table 10.2 for $n = 100$ per group ranges from 0.1195 to 0.6795. This LL indicates that our CI (provided that the sample produced $d = 0.40$) would rule out only very small population effect sizes (i.e., anything less than 0.1195).

As a brief aside, compare the CI around the effect sizes (0.1195, 0.6795) to the interval for the mean differences (0.6056, 3.3944). Taking the mean differences and dividing by the standard deviation of 5.0 does not provide exact confidence limits around the effect size. Dividing the mean difference limits by the standard deviation yields a LL of 0.1211 and an UL of 0.6789. These values are close to the confidence limits for the effect size, but are not exact. CIs around mean differences are constructed using a central *t*-distribution whereas the effect size intervals use the noncentral *t*-distribution. As noted in Chapters 2 and 3, with smaller samples, estimates based on the central *t* diverge considerably from estimates based on noncentral distributions.

The results shown in Table 10.2 indicate that an interval that is precise to 0.20 units requires a sample of 750 participants per group. That is a very large sample, but it does provide a particularly narrow range of effect sizes.

TABLE 10.2 R Code and Output for Confidence Interval around Cohen's *d* Precision Analysis

```
d_prec(d=.4, nlow=100, nhigh=2000, propn1=.5, ci=.95, by=100)
## [1] "n1 = 50, n2 = 50 d = 0.4, LL = 0.003, UL = 0.795,
precision = 0.792"
## [1] "n1 = 100, n2 = 100 d = 0.4, LL = 0.1195, UL = 0.6795,
precision = 0.56"
## [1] "n1 = 150, n2 = 150 d = 0.4, LL = 0.1711, UL = 0.6283,
precision = 0.4572"
## [1] "n1 = 200, n2 = 200 d = 0.4, LL = 0.2018, UL = 0.5977,
precision = 0.3959"
## [1] "n1 = 250, n2 = 250 d = 0.4, LL = 0.2227, UL = 0.5769,
precision = 0.3542"
## [1] "n1 = 300, n2 = 300 d = 0.4, LL = 0.2382, UL = 0.5615,
precision = 0.3233"
## [1] "n1 = 350, n2 = 350 d = 0.4, LL = 0.2502, UL = 0.5495,
precision = 0.2993"
## [1] "n1 = 400, n2 = 400 d = 0.4, LL = 0.2599, UL = 0.5398,
precision = 0.2799"
## [1] "n1 = 450, n2 = 450 d = 0.4, LL = 0.2679, UL = 0.5319,
precision = 0.264"
## [1] "n1 = 500, n2 = 500 d = 0.4, LL = 0.2747, UL = 0.5251,
precision = 0.2504"
## [1] "n1 = 550, n2 = 550 d = 0.4, LL = 0.2805, UL = 0.5193,
precision = 0.2388"
## [1] "n1 = 600, n2 = 600 d = 0.4, LL = 0.2856, UL = 0.5142,
precision = 0.2286"
## [1] "n1 = 650, n2 = 650 d = 0.4, LL = 0.2901, UL = 0.5097,
precision = 0.2196"
## [1] "n1 = 700, n2 = 700 d = 0.4, LL = 0.2941, UL = 0.5057,
precision = 0.2116"
## [1] "n1 = 750, n2 = 750 d = 0.4, LL = 0.2977, UL = 0.5022,
precision = 0.2045"
## [1] "n1 = 800, n2 = 800 d = 0.4, LL = 0.301, UL = 0.4989,
precision = 0.1979"
## [1] "n1 = 850, n2 = 850 d = 0.4, LL = 0.3039, UL = 0.496,
precision = 0.1921"
## [1] "n1 = 900, n2 = 900 d = 0.4, LL = 0.3066, UL = 0.4933,
precision = 0.1867"
## [1] "n1 = 950, n2 = 950 d = 0.4, LL = 0.3091, UL = 0.4908,
precision = 0.1817"
## [1] "n1 = 1000, n2 = 1000 d = 0.4, LL = 0.3114, UL =
0.4885, precision = 0.1771"
```

Another approach is to design for an interval that excludes a certain effect. For example, Cohen (1988) suggests $d = 0.20$ as the criterion for a small effect. Despite the reluctance expressed throughout this book to design around small, medium, and large effect size conventions, an attractive strategy for CIs is to find the sample size that yields a CI where the LL exceeds $d = 0.20$. This strategy allows for claims that at worst, the effect was small. Of course, whether $d = 0.20$ is meaningful is

another issue. A sample of $n = 200$ per group corresponds to a result where the LL of the effect size exceeds $d = 0.20$. This reflects addition of 200 participants (100 per group) over that necessary to obtain power of .80. This increase in sample size achieved a change of 0.1641 in precision (0.56–0.3959). Improving the precision by that much again (to roughly .23), requires nearly 600 participants per group, an increase of 800 participants (see sample for N1 and N2 = 600). A rough rule of thumb is that to cut error in half we need to quadruple the sample size.

Precision for a Correlation

Another form of CI around an effect size is the CI around ρ, the population correlation. Formula 10.2 defines this CI. The value z_ρ reflects the Fisher transformed correlation (see Formula 4.3). This approach uses a central distribution (the normal distribution) so it is possible to calculate the CI by hand. The part of the equation with 1 over the square root of the sample size minus 3 is termed the standard deviation of Fisher's z or sd_z. The final step in constructing this interval is to convert values back to correlation units using Formulae 10.3. This calculation reverses the Fisher's transformation (see Chapter 4 for examples discussion of the transformations). As before, I present a 95% CI. To produce other intervals simply replace $z_{.95}$ with the value of interest. For example, a 99% CI would use $z_{.01}$ whereas a 90% CI uses $z_{.90}$

$$z_\rho \pm z_{0.95} \frac{1}{\sqrt{n-3}} \tag{10.2}$$

$$\rho = \frac{e^{2z_\rho} - 1}{e^{2z_\rho} + 1} \tag{10.3}$$

Example 10.3: Confidence Limits around *r*

Chapter 4 presented an example examining the correlation between implicit attitudes and aggression where a meaningful correlation was .30. In that example, a sample of 84 participants produced power of .80. Extending this example, imagine that we wanted a correlation that was precise to .10 in either direction (precision would be .20 in this case). Table 10.3 demonstrates use of the r_prec function for completing this calculation. The format of the function is as follows:

r_prec(r, nlow, nhigh, ci, by)

The value *r* represents the correlation. nlow and nhigh define the range of sample sizes. by specifies the increase in sample size from nlow. ci determines the type of CI.

TABLE 10.3 R Code and Output for Confidence Interval around Correlation Precision Analysis

```
r_prec(r=.3, nlow=80, nhigh=400, by=20, ci=.95)
## [1] "n = 80 r = 0.3, LL = 0.0859, UL = 0.4876, precision =
0.4017"
## [1] "n = 100 r = 0.3, LL = 0.1101, UL = 0.4688, precision =
0.3587"
## [1] "n = 120 r = 0.3, LL = 0.1276, UL = 0.4548, precision =
0.3272"
## [1] "n = 140 r = 0.3, LL = 0.1411, UL = 0.4438, precision =
0.3027"
## [1] "n = 160 r = 0.3, LL = 0.1519, UL = 0.4349, precision =
0.283"
## [1] "n = 180 r = 0.3, LL = 0.1608, UL = 0.4275, precision =
0.2667"
## [1] "n = 200 r = 0.3, LL = 0.1683, UL = 0.4212, precision =
0.2529"
## [1] "n = 220 r = 0.3, LL = 0.1747, UL = 0.4158, precision =
0.2411"
## [1] "n = 240 r = 0.3, LL = 0.1802, UL = 0.411, precision =
0.2308"
## [1] "n = 260 r = 0.3, LL = 0.1851, UL = 0.4068, precision =
0.2217"
## [1] "n = 280 r = 0.3, LL = 0.1894, UL = 0.403, precision =
0.2136"
## [1] "n = 300 r = 0.3, LL = 0.1933, UL = 0.3997, precision =
0.2064"
## [1] "n = 320 r = 0.3, LL = 0.1968, UL = 0.3966, precision =
0.1998"
## [1] "n = 340 r = 0.3, LL = 0.2, UL = 0.3938, precision =
0.1938"
## [1] "n = 360 r = 0.3, LL = 0.2029, UL = 0.3912, precision =
0.1883"
## [1] "n = 380 r = 0.3, LL = 0.2056, UL = 0.3889, precision =
0.1833"
## [1] "n = 400 r = 0.3, LL = 0.2081, UL = 0.3867, precision =
0.1786"
```

The output in Table 10.3 provides the precision estimates. The desired level of precision of .20 requires a sample of over 300 participants.

An important feature of the correlation CI (as well as any CI based on effect sizes) is that the intervals are not symmetrical. Take for example the interval for $n = 80$. The interval, based on a correlation of .30, ranges from .09 to .48. Because of asymmetry of the sampling distribution, the LL is .21 units below .30 but the UL is .18 above. Example 4.1 found that a sample size of 84 produced Power = .80. Even with 100 per group, the CI in Table 10.3 is wide (.11 to .47) and imprecise. As in previous examples, conventional levels of statistical power do not translate to precise estimates.

Example 10.4: Precision for R^2

In Chapter 8, an example addressed a situation where R^2 *Model* = .467. A sample of 24 participants yielded power of .90 for R^2. The code shown in Table 10.4 demonstrates use of the R2_prec function to examines confidence limits on R^2, starting with $n = 24$ and including additional values to show a range of limits. The approaches presented here use a fixed effects approach.

TABLE 10.4 R Code and Output for R^2 Model Precision Analysis

```
R2_prec(R2=.467, nlow=24, nhigh=100, pred=3, by=4)
## [1] "n = 24 R2 = 0.467, LL = 0.0693, UL = 0.6242, precision
= 0.5549"
## [1] "n = 28 R2 = 0.467, LL = 0.1065, UL = 0.618, precision
= 0.5115"
## [1] "n = 32 R2 = 0.467, LL = 0.1365, UL = 0.6124, precision
= 0.4759"
## [1] "n = 36 R2 = 0.467, LL = 0.161, UL = 0.6074, precision
= 0.4464"
## [1] "n = 40 R2 = 0.467, LL = 0.1814, UL = 0.6029, precision
= 0.4215"
## [1] "n = 44 R2 = 0.467, LL = 0.1987, UL = 0.5989, precision
= 0.4002"
## [1] "n = 48 R2 = 0.467, LL = 0.2135, UL = 0.5952, precision
= 0.3817"
## [1] "n = 52 R2 = 0.467, LL = 0.2264, UL = 0.5918, precision
= 0.3654"
## [1] "n = 56 R2 = 0.467, LL = 0.2377, UL = 0.5887, precision
= 0.351"
## [1] "n = 60 R2 = 0.467, LL = 0.2476, UL = 0.5858, precision
= 0.3382"
## [1] "n = 64 R2 = 0.467, LL = 0.2565, UL = 0.5832, precision
= 0.3267"
## [1] "n = 68 R2 = 0.467, LL = 0.2645, UL = 0.5807, precision
= 0.3162"
## [1] "n = 72 R2 = 0.467, LL = 0.2717, UL = 0.5784, precision
= 0.3067"
## [1] "n = 76 R2 = 0.467, LL = 0.2783, UL = 0.5762, precision
= 0.2979"
## [1] "n = 80 R2 = 0.467, LL = 0.2843, UL = 0.5742, precision
= 0.2899"
## [1] "n = 84 R2 = 0.467, LL = 0.2898, UL = 0.5722, precision
= 0.2824"
## [1] "n = 88 R2 = 0.467, LL = 0.2949, UL = 0.5704, precision
= 0.2755"
## [1] "n = 92 R2 = 0.467, LL = 0.2996, UL = 0.5687, precision
= 0.2691"
## [1] "n = 96 R2 = 0.467, LL = 0.3039, UL = 0.5671, precision
= 0.2632"
## [1] "n = 100 R2 = 0.467, LL = 0.308, UL = 0.5655, precision
= 0.2575"
```

TABLE 10.5 R Code and Output for Mean Difference "Support the Null" Analysis

```
md_prec(m1=0, m2=0, s1=5, s2=5, nlow=100, nhigh=40000, propn1=.5,
ci=.95, by=100)
## [1] "n1 = 50, n2 = 50, d = 0, LL = -1.96, UL = 1.96,
precision = 3.92"
## [1] "n1 = 100, n2 = 100, d = 0, LL = -1.3859, UL = 1.3859,
precision = 2.7718"
## [1] "n1 = 150, n2 = 150, d = 0, LL = -1.1316, UL = 1.1316,
precision = 2.2632"
## [1] "n1 = 200, n2 = 200, d = 0, LL = -0.98, UL = 0.98,
precision = 1.96"
## [1] "n1 = 250, n2 = 250, d = 0, LL = -0.8765, UL = 0.8765,
precision = 1.753"
## [1] "n1 = 300, n2 = 300, d = 0, LL = -0.8002, UL = 0.8002,
precision = 1.6004"
## [1] "n1 = 350, n2 = 350, d = 0, LL = -0.7408, UL = 0.7408,
precision = 1.4816"
## [1] "n1 = 400, n2 = 400, d = 0, LL = -0.693, UL = 0.693,
precision = 1.386"
## [1] "n1 = 450, n2 = 450, d = 0, LL = -0.6533, UL = 0.6533,
precision = 1.3066"
## [1] "n1 = 500, n2 = 500, d = 0, LL = -0.6198, UL = 0.6198,
precision = 1.2396"
## [1] "n1 = 550, n2 = 550, d = 0, LL = -0.591, UL = 0.591,
precision = 1.182"
## [1] "n1 = 600, n2 = 600, d = 0, LL = -0.5658, UL = 0.5658,
precision = 1.1316"
## [1] "n1 = 650, n2 = 650, d = 0, LL = -0.5436, UL = 0.5436,
precision = 1.0872"
## [1] "n1 = 700, n2 = 700, d = 0, LL = -0.5238, UL = 0.5238,
precision = 1.0476"
## [1] "n1 = 750, n2 = 750, d = 0, LL = -0.5061, UL = 0.5061,
precision = 1.0122"
## [1] "n1 = 800, n2 = 800, d = 0, LL = -0.49, UL = 0.49,
precision = 0.98"
## [1] "n1 = 3050, n2 = 3050, d = 0, LL = -0.2509, UL = 0.2509,
precision = 0.5018"
## [1] "n1 = 4800, n2 = 4800, d = 0, LL = -0.2, UL = 0.2,
precision = 0.4"
## [1] "n1 = 8550, n2 = 8550, d = 0, LL = -0.1499, UL = 0.1499,
precision = 0.2998"
## [1] "n1 = 19200, n2 = 19200, d = 0, LL = -0.1, UL = 0.1,
precision = 0.2"
```

For information on precision for random effects, see Kelley (2008). The format of the function is as follows:

R2_prec(R2, nlow, nhigh, ci, by)

The value R2 represents R^2 for the model. nlow and nhigh define the range of sample sizes. by specifies the increase in sample size from nlow. ci determines the type of CI.

Of particular interest is the result for $n = 24$, shown in Table 10.4. In the example from Chapter 8, a sample of 24 participants gave power of .90 for R^2. However, the CI ranges from .0693 to .6242. This interval is quite wide and the LL is uninspiring. This interval suggests the variance explained by the predictors could be anywhere from roughly 7 to 62%. Doubling the sample size produces a considerably more precise interval (ranging from .2135 to .5952).

Supporting Null Hypotheses[1]

Analyses of precision provide a context for discussion of designing to support null hypotheses. Of course, "support the null hypothesis" is usually not a valid statement because the null generally refers to a statement that is rarely true (see Loftus, 1996). For example, a typical null hypothesis for a two-group comparison states that the difference between the two groups is exactly zero. In most situations, a difference of exactly zero is not plausible. More importantly, it does not usually matter if the group means are exactly equal or if they are merely very similar in the population. It is far more important to be able to determine if the means differ by enough to be practically important.[2]

Differing enough to be important or meaningful is the key to testing claims of support for the null hypothesis. For example, if we establish that a reasonable range of estimates (i.e., confidence limits) for the mean differences in the population fall below the criteria set for a meaningful difference then there is support for the conclusion that the differences between groups are likely not large enough to matter. Practically, this conclusion indicates that the differences between the groups are not deviant enough from zero to suggest a meaningfully important difference in the population.

Example 10.5: "Supporting" Null Hypotheses

The first step in this process is to determine a value that reflects the smallest meaningful difference between the groups. This is the same process as for power analyses focusing on differences. The next step is to establish the minimum precision necessary to construct a CI that excludes certain effects. Returning to the example from Chapter 3, we determined that a difference favoring a tutorial assignment over a standard assignment would have to be 2 points or more to be meaningful given the investment of instructor time for implementation. An analysis that produces a CI that falls entirely below 2.0 would suggest that the tutorial was not effective enough to improve learning meaningfully.

One approach to this question explores precision for a situation where the null is true (means exactly equal in the population). This is not a realistic expectation but it does provide some focus for the analysis. This analysis involves minor modifications to the code in Table 10.1, changing it so that the means for both groups are equal.

The precision values found in Table 10.5 are particularly useful in helping to determine sample size. First, note that at 50 participants per group, the CI excludes the meaningful difference of 2.0. This result suggests that if differences existed between the groups they likely were not big enough to be meaningful. However, note that 50 participants per group yields this result only when the sample means are equal. For a study designed to conclude that the groups do not differ, a design with 50 people per group provides enough precision only if the means differ by zero or the relationship is in the opposite direction in the sample.

Of course, this discussion fails to address how much precision we need. The most practical answer when dealing with supporting null results is often "How much can you afford?" Note that with 200 participants per group precision would be slightly less than 2 points. This level of precision means that sampled differences between means of less than a single point produce a result that supports a claim of no meaningful difference. To double precision (i.e., reducing the width of the CI by half), requires roughly 800 additional participants per group. For reference, the bottom of the table shows sample size necessary for precision of 0.5, 0.4, 0.3, and 0.2.

The basic approach outlined in this section is applicable to any of the precision analyses appearing in this chapter. Simply choose a meaningful effect size and then determine precision.

One final note on the "support the null" approach. As the example in this section highlights, ruling out small effects requires large samples. To demonstrate equivalency between groups, be clear on the resources required before beginning.

Additional Issues

In this section, I address the balance between precision and sample size. It is important to note that the present chapter scratches the surface on confidence limits on effect sizes and precision analysis. There are several outstanding resources on this topic. The MBESS package for R provides tools for many CI calculations (see also Kelley, 2007b; Kelley & Maxwell, 2003).

Precision Versus Sample Size

Examining the relationship between precision and sample size is a useful guide to determining reasonable levels of precision. For tests involving mean differences, doubling precision requires quadrupling sample size. For example, in Table 10.1, a sample of 50 participants per group produces a CI with a roughly 4-point width (approximately 0 to 4.0). An interval that is twice as precise (i.e., the width is 2.0) requires a sample of 200 per group. To obtain an interval twice as precise as found for $n = 200$ per group requires $n = 800$ per group.

The relationship is similar for precision estimates of *d* and *r*. Figure 10.1 presents precision for large effects but the same general pattern holds for small and medium effects. For *d*, precision falls below 0.5 at around $n = 275$, falls to 0.4 around $n = 400$, drops to 0.3 near $n = 700$, but does not hit 0.2 until roughly $n = 1600$. For *r*, precision is about .20 at $n = 275$, near .15 at $n = 500$, about .10 at $n = 1100$, and does not reach .05 until around $n = 4500$.

The information in Table 10.1 and Figure 10.1 should inform decisions about designing for greater precision. If you want a more precise result and can afford another 100 participants, that is a great investment when moving from $n = 100$ to $n = 200$. However, a similar investment returns very little added precision when moving from $n = 2000$ to $n = 2100$.

Summary

Precision analysis addresses the sample size required to produce a CI with a particular width. Whereas power analysis addresses sample sizes required for detecting a nonzero effect, precision analyses are relevant to questions of accurately estimating population parameters. Designing for considerable precision often requires larger sample sizes than for power analysis. This chapter presented precision analyses for mean differences and effect sizes including *r*, *d*, and R^2. For most analyses, the primary information required is the desired level of precision.

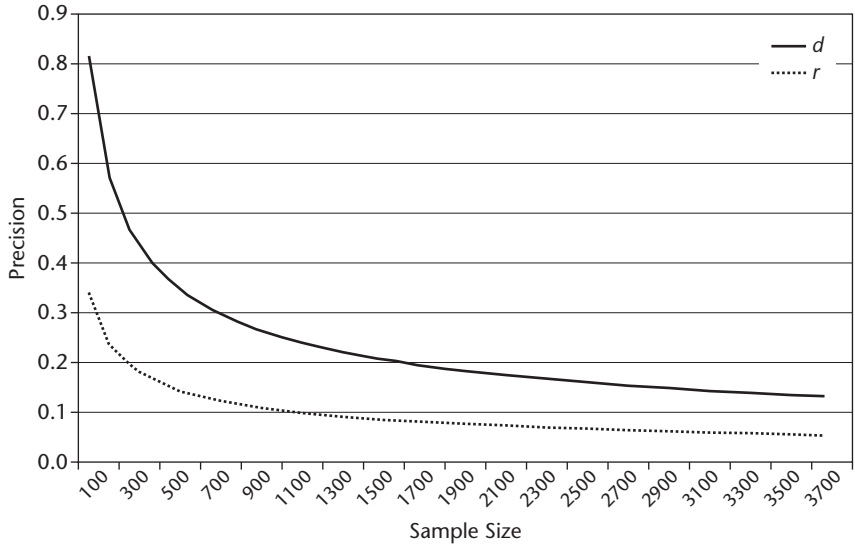

FIGURE 10.1 Precision and Sample Size.

Notes

1. This section assumes a frequentist perspective. There are Bayesian approaches relevant to "supporting" null results as well.
2. Recent work on equivalence testing (Lakens, 2017b) power provide results that are functionally equivalent to those discussed in this section. The TOSTER package is an excellent resource for those and other approaches.

11

ADDITIONAL ISSUES AND RESOURCES

This chapter presents a variety of topics including reporting power analyses, testing assumptions, converting between effect size estimates, additional resources for power, sources for learning about analyses not covered in this text, and how to deal with them, and improving power without increasing sample size.

Accessing the Analysis Code

Interested readers can investigate all of the functions used in the text by opening the individual functions in R. Functions can be found at https://github.com/chrisaberson/pwr2ppl (see the R directory). The name of each function corresponds to the command used in the text.

Using Loops to Get Power for a Range of Values

Most of the functions in pwr2ppl provide power for a single sample size. An additional piece of code using loops provides power across a range of sample sizes. This approach can be used with just about any of the functions in the package. The approach detailed below gives power estimates for the three predictor multiple regression example in Chapter 8.

```
for (i in seq(100, 200, 10))
{MRC(ry1=.40, ry2=.40, ry3=–.40, r12=–.15, r13=–.60, r23=.25, n=i)}
```

The first part, for (i in seq), is what is a for-loop command. The basic idea is for each value in the sequence that follows, perform the command wrapped in the {}. The three numbers that follow (100, 200, 10) feed values of 100 to 200 in

increments of 10 to the MRC command below. This produces analyses for sample sizes of 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, and 200. Inside the MRC command, values are as before, except the *n* is now set equal to i (corresponding to the "i in seq" part of the command).

How to Report Power Analyses

After conducting a power analysis, it is important to report the analysis accurately and completely. Power analysis is a design issue, so discussions of power go in the Method section of an American Psychological Association (APA) style paper. As Wilkinson and the Task Force for Statistical Inference noted

> [b]ecause power computations are most meaningful when done before data are collected and examined, it is important to show how effect size estimates have been derived from previous research and theory in order to dispel suspicions that they might have been taken from data used in the study or, even worse, constructed to justify a particular sample size.
>
> *(1999, p. 596)*

Similarly, the *APA Style Manual* instructs authors to "give the intended size of the same and number of individuals meant to be in each condition if separate conditions were used … [s]tate how this intended sample size was determined (e.g., analysis of power or precision)" (2010, pp. 30–31).

Based on these statements, there is a clear directive that researchers should report in detail how they arrived at decisions about a meaningfully sized effect and all the statistical values used or assumed in estimation. However, it appears that these recommendations are not widely followed. As noted in Chapter 1, recent changes to editorial policies at several outlets suggest reporting power analyses is now a point of emphasis.

Suggested Guidelines for Reporting Statistical Power

1. Report power analyses for all focal hypotheses.
2. Justify each value used in the analysis. If you design around an effect size of $d = 0.50$, explain why you selected that value. Do not simply note that $d = 0.50$ corresponds to a medium-sized effect.
3. Design to detect the smallest effect size of interest. Justify why that effect size is the smallest of interest.
4. When using complex designs, discuss all considerations that influence power. For example, in multiple regression, when designing to detect a particular R^2 value, discuss all the correlations in the design. Justify your decisions.
5. If using software approaches, cite the source of power calculations.

Example 11.1: Reporting a Power Analysis for a Chi-Square Analysis

To determine sample size requirements for the present study, I first estimated a baseline value for rental availability based on Page (1999) who found that renters indicated to 76% of those in the control group that the property was available. Next, I determined that a 20% difference between groups (i.e., the HIV group hearing "available" 56% of the time), constituted a meaningful difference between the two groups. In determining how large a difference would be meaningful, I note that previous work detected larger differences (36%). A 20% difference allows for detection of smaller effects than found in the original study while allowing for detection of considerable levels of discrimination. These values correspond to $\Phi = .21$. The techniques outlined in Aberson (2019) found a sample of 180 participants would yield power of .80 for detecting this effect.

Example 11.2: Reporting a Power Analysis for Repeated Measures ANOVA

To determine sample size requirements for the present study, we examined use of the stereotype negation procedure in other samples. These techniques produced standard deviations of approximately 0.40 with raw score changes of +0.25 to +0.40 (meaning more positive attitudes) for pre to post improvement and gradual increases thereafter. Based on this information, we judged +0.25 as the minimum value for a practically important pre-post change with smaller changes expected from the posttest to 2-hour measurement and from the 2-hour to 6-hour measure. Previous work reported test–retest reliability at .50. However, correlations between measures often decay over time so we set pre-2-hour and 2 hour–6 hour correlations at .30 and the pre-6-hour correlation at .15. We also expected standard deviations to increase slightly over time so we set the posttest standard deviation at 0.50, the 2-hour SD at 0.60, and the 6-hour SD at 0.70. Changes to the standard deviations and correlations produce more conservative power estimates than would use of the initial estimates across each measurement period. These parameters reflect an omnibus effect size of $\eta^2 = .14$. Since the expected pattern of correlations suggest issues with sphericity, we adjusted power estimates for a test using the Greenhouse–Geisser adjustment. Based on these values, power analyses following the procedures outlined by Aberson (2019) found that a sample of 29 participants produced adequate power for the omnibus test (.81).

Reporting Power if Not Addressed A Priori

It is not uncommon for researchers to collect data without a clear power analysis to guide sample size. In cases like this it is still useful to report power for the obtained sample size. In such reports, clearly note that data collection reflected

convenience, cost limitations, or whatever factors drove your approach. Then note, given your sample size, what sort of power your sample affords.

For example, if you conducted a simple between subjects treatment-control comparison, a reasonable power analysis might read as follows. Data collection reflected the maximum number of participants that could be obtained over the course of a single semester. A sample of 45 participants per group (90 overall), yields 80% power to detect effects of $d = 0.60$, 90% power to detect $d = 0.70$, and 95% power to detect $d = 0.77$.

This approach is not ideal but it does allow readers a clear understanding of the sort of effects that might reasonably be detected given your sample size. For this approach, a larger sample size leads to more convincing power statements as you do not have the opportunity to make a case for why you designed around a particularly effect size.

Statistical Test Assumptions

Assumptions are mentioned throughout the text with regard to specific tests. Regardless of study design, most statistical procedures are most accurate when data meet test assumptions. More often than not, meeting assumptions yields more power for data analyses. This is especially important when dealing with relatively small samples. Pay careful attention to issues such as data cleaning and assumptions prior to analyses as these can impact power considerably.

Effect Size Conversion Formulae

Occasionally, it is useful to convert between effect size estimates. Formulae below address conversions between several major estimates.

Eta squared to d

The values p_1 and p_2 are the proportion of participants in each group. For equal sample sizes ($p_1 = p_{2=}.50$), the numerator simplifies to $4\eta^2_{partial}$. Some sources present this conversion formula with $4\eta^2_{partial}$ in the numerator, but that approach is only appropriate for equal sample sizes whereas the proportional values presented in Formula 11.1 are applicable to equal and unequal samples.

$$d = \sqrt{\frac{\eta^2_{partial} \,/\, p_1 p_2}{1 - \eta^2_{partial}}} \tag{11.1}$$

d *to Eta squared*

For equal sample sizes, the denominator in Formula 11.2 simplifies to $d^2 + 4$. Some versions of this formula include $d^2 + 4$ in the denominator but that is appropriate only when sample sizes are equal across groups.

$$\eta^2 = \frac{d^2}{d^2 + \frac{1}{p_1 p_2}} \tag{11.2}$$

Correlation to d

This approach applies only to between group designs. I use ρ to note the population correlation in Formulae 11.3 and 11.4. When dealing with samples simply substitute r for ρ

$$d = \sqrt{\frac{\rho^2 / p_1 p_2}{(1 - \rho^2)}} \tag{11.3}$$

d *to Correlation*

Again, this is applicable only to between group designs.

$$\rho = \sqrt{\frac{d^2}{d^2 + \frac{1}{p_1 p_2}}} \tag{11.4}$$

General (Free) Resources for Power and Related Topics

Two excellent programs for power analysis are available as freeware. The first is G*Power 3 (see www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/). The second resource is a set of applets called PiFace (see www.math.uiowa.edu/~rlenth/Power/). Both programs are easy to use for simple analyses. G*Power's authors provide papers that detail most of the procedures (Faul, Erdfelder, Lang, & Buchner, 2007; Faul, Erdfelder, Buchner, & Lang, 2009). For both programs, users must make sure to understand completely what values the program requires for input. Both programs use language that may be unfamiliar to some users, particularly for complex analyses. When using one of these programs, I strongly recommend checking results against those produced by another program or testing an example where the correct answer is known. If you misunderstand the required input, power analysis is meaningless.

The Web Interface for Statistics Education (wise.cgu.edu) provides an interactive power tutorial that is an outstanding resource for learning about power analysis (note: I co-wrote the tutorial). The website also provides tutorials on other relevant topics like the Central Limit Theorem and hypothesis testing as well as an easy-to-use spreadsheet for calculating distribution probabilities. Another tool useful for hand calculations is the Noncentral Distribution Calculator (see www.statpower.net). Finally, there are several excellent protocols in Stata, SAS, and R for simple power analyses. See https://stats.idre.ucla.edu/other/dae/ for a summary of these approaches.

Resources for Additional Analyses

There are many analyses not covered in the present text and analyses where presentation was limited. For these topics, the listings below provide helpful resources and references.

Confidence Intervals around Effect Sizes

A number of tools exist for confidence intervals (CI) around effect sizes. Michael Smithson's text (2003) on CI is particularly useful. His web site at www.michaelsmithson.online/stats/CIstuff/CI.html also includes resources written for SPSS, SAS, SPlus, and R. Ken Kelley's MBESS web site contains analysis packages for confidence limits around most parameters (nd.edu/~kkelley/site/MBESS.html; see also Kelley, 2007a; 2007b; 2008).

Another outstanding tool is Exploratory Software for Confidence Intervals (ESCI). ESCI provides modules for exploring CI and noncentral distributions. I used ESCI to create several of the figures in Chapters 1 and 3. See https://thenewstatistics.com/itns/esci/ for materials.

Mediation Power

For power of mediated effects, Fritz and MacKinnon (2007) provide an overview and power tables. Approaches for using Monte Carlo methods to test mediation models and address power for both parallel and serial mediation designs exist at http://marlab.org/power_mediation/ (Schoemann, Boulton, & Short, 2017).

Structural Equations Modeling Power

Several approaches are available for power analyses for structural equations modeling. MacCallum, Browne, and Sugawara (1996) present tables for addressing close, not close, and exact fit. Also see Satorra and Saris (1985) and the May 2007 issue of *Personality and Individual Differences* that is devoted to structural equations modeling. More recent work on Monte Carlo provides additional guidance (Wolf, Harrington, Clark, & Miller, 2013).

Multilevel Modeling Power

An outstanding resource for multilevel modeling modeling (also known as hierarchical linear modeling) is the Optimal Design program and accompanying manual (Spybrook, Raudenbusch, Liu, Congdon, & Martínez, 2008). Both are available from sitemaker.umich.edu/group-based/optimal_design_software. Monte Carlo approaches provide addition guidance (Lane & Hennes, 2018).

Improving Power without Increasing Sample Size or Cost

The major focus of this text is statistical approaches to power analysis where the remedy for low power is usually the addition of more participants. However, several methodological approaches increase power without increasing sample size. These are great options to consider before adding participants. There are obvious benefits to increasing power without adding costs associated with larger samples. Several of these suggestions receive a more thorough consideration in Lipsey (1990).

Stronger experimental manipulations increase effect size. Stronger manipulations result from stronger treatments, weaker controls, or both. As an example of this, a few years back I conducted a series of studies that manipulated the qualifications and ethnicity of potential job applicants. Manipulations consisted of a cover page attached to a questionnaire that summarized the applicant's qualifications and presented a one paragraph personal statement that varied ethnicity. Later I supervised a project where a student modified the approach by creating files for each applicant that included a resume on nice paper and a photograph of the applicant. The student's study produced a considerably larger effect size than the earlier studies, likely because of the stronger and more engaging manipulation.

Another option is assigning more participants to cheaper conditions and fewer to more expensive conditions, or sampling relatively more participants from cheaper or easier-to-obtain groups. When sampling from existing groups this strategy can be of great use. Chapter 3 presented an example comparing gay men from the community to heterosexual men from campus. The gay men received monetary compensation for their time but the campus sample participated for course credit. The campus sample could have been increased considerably with minimal cost, resulting in substantial increases in power.

Simplifying research designs reduces sample size requirements as well. Researchers often strive to answer so many questions that the design becomes overwhelming. For example, imagine a $2 \times 2 \times 2$ design where power analyses suggest 25 participants per cell. Cutting out a factor makes this a 2×2 design and likely reduces total sample size requirements considerably. On a similar note, researchers should always ask whether all factor levels are necessary in designs with more than two levels.

Within subjects designs usually produce considerably more power than between subjects approaches. Although within subjects approaches are not always possible, they are likely underutilized. Researchers with concerns about carryover effects might evaluate carryover by pretesting using within subjects approaches.

Several chapters included discussions of reliability. Poor or even mediocre reliability reduces observed effect sizes considerably. Insist on the most reliable measures possible.

Finally, I want to close with something a colleague once said about power and research design. I consulted briefly on a project and asked my colleague if he had conducted a power analysis. He laughed and said, "that stuff is for people who don't understand research design." I do not agree entirely, but it is clear that good research design substantially improves statistical power.

REFERENCES

Aberson, C. L. (2007). Diversity, merit, fairness, and discrimination beliefs as predictors of support for affirmative action policy actions. *Journal of Applied Social Psychology, 37*, 2451–2474. doi:http://dx.doi.org/10.1111/j.1559-1816.2007.00266.x.

Aberson, C. L. (2019). *Applied power analysis for the behavioral sciences* (2nd edition). Routledge: New York.

Aberson, C. L., Berger, D. E., Healy, M. R., & Romero, V. L. (2002). An interactive tutorial for teaching statistical power. *Journal of Statistics Education, 10*, 3. Retrieved from ww2.amstat.org/publications/jse/v10n3/aberson.html.

Aberson, C. L., & Gaffney, A. M. (2009). An integrated threat model of implicit and explicit attitudes. *European Journal of Social Psychology, 39*, 808–830. doi:http://dx.doi.org/10.1002/ejsp.582.

Aberson, C. L., Healy, M. R., & Romero, V. L. (2000). Ingroup bias and self-esteem: A meta-analysis. *Personality and Social Psychology Review, 4*, 157–173. doi:http://dx.doi.org/10.1207/S15327957PSPR0402_04.

Aguinis, H. (2004). *Regression analysis for categorical moderators*. New York: Guilford Press.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology, 90*, 94–107. doi:http://dx.doi.org/10.1037/0021-9010.90.1.94.

Aguinis, H., Petersen, S. A., & Pierce, C. A. (1999). Appraisal of the homogeneity of error variance assumption and alternatives to multiple regression for estimating moderating effects of categorical variables. *Organizational Research Methods, 2*, 315–339. doi:http://dx.doi.org/10.1177/109442819924001.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

American Psychological Association (n.d.a.). New editor spotlight. Retrieved from www.apa.org/pubs/highlights/editor-spotlight/psp-asc-kitayama.aspx.

American Psychological Association (n.d.b.). New editor spotlight. Retrieved from www.apa.org/pubs/highlights/editor-spotlight/emo-pietromonaco.aspx.

Anderson, C. A., & Bushman, B. J. (2002). Human aggression. *Annual Review of Psychology, 53*, 27–51. doi:http://dx.doi.org/10.1146/annurev.psych.53.100901.135231.

Ayduk, O., Gyurak, A., & Luerssen, A. (2008). Individual differences in the rejection-aggression link in the hot sauce paradigm: The case of rejection sensitivity. *Journal of Experimental Social Psychology, 44*, 775–782. doi:http://dx.doi.org/10.1016/j.jesp.2007.07.004.

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, M. (2016). Researchers' intuitions about power in psychological research. *Psychological Science, 27*, 1069–1077. doi:10.1177/0956797616647519.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10*, 389–96. doi:http://dx.doi.org/10.1146/annurev.psych.53.100901.135231

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*, 407–425. doi: 10.1037/a0021524.

Benjamin, D. J., Berger, J., Johannesson, M., Nosek, B. A., Wagenmakers, E., … Johnson, V. (2017, July 22). Redefine statistical significance. doi:10.17605/OSF.IO/MKY9J.

Blanca, M., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit? *Behavior Research Methods, 50*, 937–962. doi:10.3758/s13428-017-0918-2.

Brown, J., & Hale, M. S. (1992). The power of statistical studies in consultation-liaison psychiatry. *Psychosomatics: Journal of Consultation Liaison Psychiatry, 33*, 437–443.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 365–376. doi:http://dx.doi.org/10.1038/nrn3475.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2017, October 2). Correcting for bias in psychology: A comparison of meta-analytic methods. Retrieved from psyarxiv.com/9h3nu.

Cashen, L. H., & Geiger, S. W. (2004). Statistical power and the testing of null hypotheses: A review of contemporary management research and recommendations for future studies. *Organizational Research Methods, 7*, 151–167.

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

Cohen, J. (1984). The cost of dichotomization. *Applied Psychological Measurement, 7*, 249–253.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1994). How the power of MANOVA can both increase and decrease as a function of the interrelations among the dependent variables. *Psychological Bulletin, 115*, 465–474.

Cumming, G., & Finch, S. (2005). Inference by eye. Confidence intervals and how to read pictures of data. *American Psychologist, 60*, 170–180.

Davis, J. R., & Henry, P. J. (2008, February). *The culture of the lab: Influences of the college setting on social psychology's view of the nature of prejudice*. Poster session presented at the annual meeting of the Society for Personality and Social Psychology.

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's *t*-test instead of Students *t*-test. *International Review of Social Psychology, 30*, 92–101. doi:10.5334/irsp.82.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149–1160. doi.org/10.3758/BRM.41.4.1149.

Faul, F., Erdfelder, E., Lang, A. L., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Field, A. (1998). A bluffer's guide to … sphericity. *Newsletter of the Mathematical, Statistical and Computing section of the British Psychological Society, 6*, 13–22.

Finch, S., Thomason, N., & Cumming, G. (2002). Past and future American Psychological Association guidelines for statistical practice. *Theory and Psychology, 12*, 825–853.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron, 1*, 3–32.

Fisher, R. A. (1938). Presidential address, Indian statistical conference. *Sankhyã, 4*, 14–17.

Fitzsimons, G. (2008). A death to dichotomizing. *Journal of Consumer Research, 35*, 5–8.

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE, 9*: e109019. https://doi.org/10.1371/journal.pone.0109019.

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science, 18*, 233–239.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*, 17–41.

Greitemeyer, T. (2009). Effects of songs with prosocial lyrics on prosocial thoughts, affect, and behavior. *Journal of Experimental Social Psychology, 45*, 186.

Harrison, D. A., Kravitz, D. A., Mayer, D. M., Leslie, L. M., & Lev-Arey, D. (2006). Understanding attitudes toward affirmative action programs in employment: Summary and meta-analysis of 35 years of research. *Journal of Applied Psychology, 91*, 1013–1036.

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York: Guilford Publications.

Hittner, J. B., May, K., & Silver, N. C. (2003). A Monte Carlo evaluation of tests for comparing dependent correlations. *The Journal of General Psychology, 130*, 149–168.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician, 55*, 19–24.

Hsieh, F. Y., Bloch, D. A., & Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression. Statistics in Medicine, *17*, 1623–1634. doi:10.1002/(SICI)1097-0258(19980730)17:14<1623::AID-SIM871>3.0.CO;2-S.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8*, 1, 3–7.

Hunter, J. E., & Schmidt, F. L. (1990). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology, 75*, 334–349.

Hunter, J. E., & Schmidt, F. L. (1994). Correcting for sources of artificial variation across studies. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 323–336). New York: Russell Sage.

Jennions, M. D., & Møller, A. P. (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology, 14*, 438–445.

Kawakami, K., Dovidio, J. F., Moll, J., Herrasen, S., & Russin, A. (2000). Just say no (to stereotyping): Effects of training in the negation of Stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology, 78*, 871–888.

Kelley, K. (2007a). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software, 20*, 1–24.

Kelley, K. (2007b). Methods for the behavioral, educational, and social science: An R package. *Behavior Research Methods, 39*, 979–984.

Kelley, K. (2008). Sample size planning for the squared multiple correlation coefficient: Accuracy in parameter estimation via narrow confidence intervals. *Multivariate Behavioral Research, 43*, 524–555.

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods, 8*, 305–321.

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods, 11*, 363–385.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Keselman, H. J., Othman, A., Wilcox, R., & Fradette, K. (2004). The new and improvement two-sample t tests. *Psychological Science, 15*, 47–51.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Kosciulek, J. F., & Szymanski, E. M. (1993). Statistical power analysis of rehabilitation counseling research. *Rehabilitation Counseling Bulletin, 36*, 212–219.

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects: Statistical power analysis in research*. Newbury Park, CA: Sage.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses: Sequential analyses. *European Journal of Social Psychology, 44*, 701–710. doi:10.1002/ejsp.2023.

Lakens, D. (2017a, May 11). How a power analysis implicitly reveals the smallest effect size you care about. Retrieved from http://daniellakens.blogspot.com/2017/05/how-power-analysis-implicitly-reveals.html.

Lakens, D. (2017b). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*, 355–362. doi:http://dx.doi.org/10.1177/1948550617697177.

Lakens, D., & Evers, E. R. K. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science, 9*, 278–292. doi:10.1177/1745691614528520.

Lane, S. P., & Hennes, E. P. (2018). Estimating sample size for multilevel relationship research. *Journal of Social and Personal Relationships, 35*, 7–31. doi:10.1177/0265407517710342.

Lenth, R. V. (2000, August). *Two sample size practices that I don't recommend*. Paper presented at the Joint Statistical Meeting, Indianapolis, IN.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician, 55*, 187–193.

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science, 26*, 1827–1832. doi:http://dx.doi.org/10.1177/0956797615616374.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5*, 161–171.

Lucas, R. E., & Donnellan, M. B. (2013). Improving the replicability and reproducibility of research published in the journal of research in personality. *Journal of Research in Personality, 47*, 453–454. doi:http://dx.doi.org/10.1016/j.jrp.2013.05.002.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods. 1*, 130–149.

Maddock, J. E., & Rossi, J. S. (2001). Statistical power of articles published in three health-psychology related journals. *Health Psychology, 20*, 76–78.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*, 147–163.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59*, 537–563.

McCartney, K., & Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. *Child Development, 71*, 173–180.

Menard, S. (2009). *Applied logistic regression analysis*. Sage.

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.

Nakagawa, S., & Foster, T. M. (2004). The case against retrospective statistical power analyses with and introduction to power analysis. *Acta Ethologica, 7*, 103–108.

Nature Publishing Group (2017). Announcement: Towards greater reproducibility for life-sciences research in Nature. *Nature, 546*, 8. doi:10.1038/546008a.

Nickerson, R. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241–301. doi:10.1037/1082-989X.5.2.241.

Onwuegbuzie, A. J., & Leech, N. L. (2004). Post-hoc power: A concept whose time has come. *Understanding Statistics, 3*, 201–230. doi:10.1207/s15328031us0304_1.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. (2015). *Science, 349*, 1–8. doi:10.1126/science.aac4716.

Page, S. (1999). Accommodating persons with AIDS: Acceptance and rejection in rental situations. *Journal of Applied Social Psychology*, 29, 261–270.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology, 75*, 811–832.

Prendergast, M. L., Podus, D., Chang, E., & Urada, D. (2002). The effectiveness of drag abuse treatment: A meta-analysis of comparison group studies. *Drug and Alcohol Dependence, 67*, 53–72.

Revelle, W. (2018). psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version=1.8.4.

Rogers, W. M. (2002). Theoretical and mathematical constraints of interactive regression models. *Organizational Research Methods, 5*, 212–230.

Rosenthal, R. (1979). The "file-drawer problem" and tolerance for null results. *Psychological Bulletin, 86*, 638–641.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*, 166–169.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58*, 646–656.

Satorra, A., & Saris, W. E. (1985). The power of the likelihood ratio test in covariance structure analysis. *Psychometrika, 50*, 83–90.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17*, 551–566. doi:10.1037/a0029487.

Schimmack, U. (2016). The Replicability-Index: Quantifying Statistical Research Integrity. https://wordpress.com/post/replication-index.wordpress.com/920.

Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining power and sample size for simple and complex mediation models. *Social Psychological and Personality Science, 8*, 379–386.

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

Siegel, S., & Castellan, J. N. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used. *Journal of Applied Psychology, 72*, 146–148.

Silver, N. C., Hittner, J. B., & May, K. (2004). Testing dependent correlations with nonoverlapping variables: A Monte Carlo simulation. *Journal of Experimental Education, 73*, 53–69.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366. doi:10.1177/0956797611417632.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*, 559–569. doi:http://dx.doi.org/10.1177/0956797614567341.

Smithson, M. J. (2003). *Confidence intervals*. Thousand Oaks, CA: Sage.

Spybrook, J., Raudenbusch, S. W., Liu, X-f., Congdon, R., & Martínez, A. (2008). Optimal design for longitudinal and multilevel research: Documentation for the "Optimal Design" software. Downloaded from http://sitemaker.umich.edu/group-based/optimal_design_software on June 20, 2008.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245–251.

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

Tabachnick, B. G., & Fidell, L. S. (2007a). *Using multivariate statistics* (5th ed.). Boston: Allyn and Bacon.

Tabachnick, B. G., & Fidell, L. S. (2007b). *Experimental designs using ANOVA*. Belmont, CA: Duxbury.

Thompson, B. (1998). Statistical significance and effect size reporting: Portrait of a possible future. *Research in the Schools, 5*, 33–38.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*, 24–31.

Tropp, L. R., & Pettigrew, T. F. (2005). Differential relationships between intergroup contact and affective and cognitive indicators of prejudice. *Personality and Social Psychology Bulletin, 31*, 1145–1158.

Unkelbach, C. (2016). Increasing replicability. *Social Psychology, 47*, 1–3. doi:http://dx.doi.org/10.1027/1864-9335/a000270.

Van Laar, C., Levin, S., Sinclair, S., & Sidanius, J. (2005). The effect of university roommate contact on ethnic attitudes and behavior. *Journal of Experimental Social Psychology, 41*, 329–345.

Vazire, S. (2016). Editorial. *Social Psychological and Personality Science,* 7, 3–7. doi:http://dx.doi.org/10.1177/1948550615603955.

West, R. F. (1985). A power analytic investigation of research in adult education: 1970–1982. *Adult Education Quarterly, 35*, 131–141. doi:https://doi.org/10.1177/0001848185035003002.

Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods, 8*, 254–274.

Wilcox, R. R., & Tian, T. (2008). Comparing dependent correlations. *Journal of General Psychology, 135*, 105–112.

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

Williams, E. J. (1959). The comparison of regression variables. *Journal of the Royal Statistical Society, Series B, 21*, 396–399.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73*, 913–934. doi:http://dx.doi.org/10.1177/0013164413495237.

INDEX

Page numbers in **bold** denote tables, those in *italics* denote figures.